View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Apollo

Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity

Danish Saleheen[1,2]†*, Pradeep Natarajan[3,4]†, Irina Armean[4,5], Wei Zhao[1], Asif Rasheed[2], Sumeet Khetarpal[6], Hong-Hee Won[7], Konrad J. Karczewski[4,5], Anne H. O'Donnell-Luria[4,5,8], Kaitlin E. Samocha[4,5], Namrata Gupta[4], Mozzam Zaidi[2], Maria Samuel[2], Atif Imran[2], Shahid Abbas[9], Faisal Majeed[2], Madiha Ishaq[2], Saba Akhtar[2], Kevin Trindade[6], Megan Mucksavage[6], Nadeem Qamar[10], Khan Shah Zaman[10], Zia Yaqoob[10], Tahir Saghir[10], Syed Nadeem Hasan Rizvi[10], Anis Memon[10], Nadeem Hayyat Mallick[11], Mohammad Ishaq[12], Syed Zahed Rasheed[12], Fazal-ur-Rehman Memon[13], Khalid Mahmood[14], Naveeduddin Ahmed[15], Ron Do[16,17], Ronald M. Krauss[18], Daniel G. MacArthur[4,5], Stacey Gabriel[4], Eric S. Lander[4], Mark J. Daly[4,5], Philippe Frossard[2]†, John Danesh[19,20]†, Daniel J. Rader[6,21]†, Sekar Kathiresan[3,4]†*

†Contributed equally

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

[2] Center for Non-Communicable Diseases, Karachi, Pakistan

[3] Center for Human Genetic Research and Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA

[4] Broad Institute of Harvard and MIT, Cambridge, MA, USA

[5] Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA

[6] Division of Translational Medicine and Human Genetics, Department of Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA

[7] Samsung Advanced Institute for Health Sciences and Technology (SAIHST), Sungkyunkwan University, Samsung Medical Center, Seoul, Korea

[8] Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA

[9] Faisalabad Institute of Cardiology, Faisalabad, Pakistan

[10] National Institute of Cardiovascular Disorders, Karachi, Pakistan

[11] Punjab Institute of Cardiology, Lahore, Pakistan

[12] Karachi Institute of Heart Diseases, Karachi, Pakistan

[13] Red Crescent Institute of Cardiology, Hyderabad, Pakistan

[14] The Civil Hospital, Karachi, Pakistan

[15] Liaquat National Hospital, Karachi, Pakistan

[16] Department of Genetics and Genomic Sciences, Mount Sinai Medical Center, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[17] The Charles Bronfman Institute of Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[18] Children's Hospital Oakland Research Institute, Oakland, CA, USA

[19] Department of Public Health and Primary Care, University of Cambridge, UK

[20] Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

[21] Department of Human Genetics, University of Pennsylvania, USA

*Corresponding authors:

Danish Saleheen, MBBS, PhD
Department of Biostatistics and Epidemiology
University of Pennsylvania
11-134 Translational Research Center
3400 Civic Center Boulevard
Philadelphia, PA 19104
Tel: 215-573-6323
Fax: 215-573-2094
Email: saleheen@mail.med.upenn.edu

Sekar Kathiresan, MD
Broad Institute and Massachusetts General Hospital
CPZN 5.830
185 Cambridge Street
Boston, MA 02114
Tel: 617-643-6120
Email: sekar@broadinstitute.org

Word Count, Summary Paragraph: 313

Word Count, Main Text: 3,123

Summary Paragraph

A major goal of biomedicine is to understand the function of every gene in the human genome.[1] Loss-of-function (LoF) mutations can disrupt both copies of a given gene in humans and phenotypic analysis of such ‘human knockouts’ can provide insight into gene function. To date, comprehensive analysis of genes knocked out in humans has been limited by the fact that LoF mutations are infrequent in the general population and so, observing an individual homozygous LoF for a given gene is exceedingly rare.[2,3] However, consanguineous unions are more likely to result in offspring who carry LoF mutations in a homozygous state. In Pakistan, consanguinity rates are notably high.[4] Here, in order to understand consequences of complete gene disruption in humans, we sequenced the protein-coding regions of 10,503 adult participants living in Pakistan, identified individuals carrying predicted LoF (pLoF) mutations in the homozygous state, and performed phenotypic analysis involving >200 traits. We enumerated 49,138 rare (<1 % minor allele frequency) pLoF mutations. These pLoF mutations are predicted to knock out 1,317 genes in at least one participant. Homozygosity for pLoF mutations at *PLAG27* was associated with absent enzymatic activity of soluble lipoprotein-associated phospholipase A2; at *CYP2F1*, with higher plasma interleukin-8 concentrations; at *TREH*, with lower concentrations of apoB-containing lipoprotein subfractions; at either *A3GALT2* or *NRG4*, with markedly reduced plasma insulin C-peptide concentrations; and at *SLC9A3R1*, with mediators of calcium and phosphate signaling. Finally, *APOC3* is a gene which regulates metabolism of plasma triglyceride-rich lipoproteins and where heterozygous deficiency confers resistance to coronary heart disease.[5,6] In Pakistan, we now observe *APOC3* homozygous pLoF carriers; we recalled these knockout humans and

challenged with an oral fat load. Compared with wild-type family members, *APOC3* knockouts displayed marked blunting of the usual post-prandial rise in plasma triglycerides. Overall, these observations provide a roadmap for a 'human knockout project', a systematic effort to understand the phenotypic consequences of complete disruption of genes in humans.

Main Text

We studied adult participants in the Pakistan Risk of Myocardial Infarction Study (PROMIS) designed to understand the determinants of cardiometabolic diseases in South Asians.[7] Consanguineous marriages have been common in this region of South Asia for many generations.[8] In PROMIS, 39.0% of participants reported that their parents were cousins and 39.8% reported themselves being married to a cousin. An expectation from consanguinity is long regions of autozygosity, defined as homozygous loci identical by descent.[9] Using genome-wide genotyping data available in 18,541 PROMIS participants, we quantified the length of runs of homozygosity, defined as homozygous segments at least 1.5 megabases long. We compared the lengths of runs of homozygosity among PROMIS participants with those seen in other populations from the International HapMap3 Project. Median length of genome-wide homozygosity among PROMIS participants was 6-7 times higher than participants of European (CEU, TSI) ($P = 3.6 \times 10^{-37}$), East Asian (CHB, JPT, CHD) ($P = 5.4 \times 10^{-48}$) and African ancestries (YRI, MKK) ($P = 1.3 \times 10^{-40}$), respectively (**Supplementary Figure 1**).

In order to identify individuals who are homozygous for predicted loss-of-function (pLoF) mutations (i.e., nonsense, frameshift, or canonical splice-site mutations predicted to inactivate a gene), we performed whole exome sequencing in 10,503 PROMIS participants (**Table 1**) with genetic ancestry similar to the overall cohort. Across all participants, 1,639,223 exonic and splice-site sequence variants in 19,026 autosomal genes passed quality control metrics. Of these, 57,137 mutations across 14,345 autosomal genes were annotated as pLoF.

To increase the probability that mutations annotated as pLoF by automated algorithms are *bona fide*, we removed nonsense and frameshift mutations occurring within the last 5% of the transcript and within exons flanked by non-canonical splice sites, splice site mutations at small (<15 bp) introns, at non-canonical splice sites, and where the purported pLoF allele is observed across primates. Common pLoF alleles are less likely to exert strong functional effects as they are less constrained by purifying selection; thus, we define pLoF mutations in the rest of the manuscript as variants with a minor allele frequency (MAF) of < 1% and passing the aforementioned bioinformatic filters. Applying these criteria, we generated a set of 49,138 pLoF mutations across 13,074 autosomal genes.[10] The site-frequency spectrum for these pLoF mutations revealed that the majority was seen only in one or a few individuals (**Supplementary Figure 2**).

Across all 10,503 PROMIS participants, both copies of 1,317 distinct genes were predicted to be inactivated due to pLoF mutations. A full listing of all 1,317 genes knocked out, the number of knockout participants for each gene, and the specific pLoF mutation(s) are provided in **Supplementary Table 1**. 891 (67.7 %) of the genes were knocked out only in one participant (**Fig. 1a**). Nearly 1 in 5 sequenced participants (1,843 individuals, 17.5 %) had at least one gene knocked by a homozygous pLoF mutation. 1,504 of these 1,843 individuals (81.6 %) were homozygous pLoF carriers for just one gene, but a minority of participants were knockouts for more than one gene and one participant had six genes with homozygous pLoF genotypes.

We compared the coefficient of inbreeding (F coefficient) in PROMIS participants with that of 15,249 individuals from outbred populations of European or

African American ancestry. The F coefficient estimates the excess homozygosity compared with an estimated outbred ancestor. PROMIS participants had a 4-fold higher median inbreeding coefficient compared to outbred populations (0.016 v 0.0041; $P < 2 \times 10^{-16}$) (**Fig. 1b**). Additionally, those in PROMIS who reported that their parents were closely related had even higher median inbreeding coefficients than those who did not (0.023 v 0.013; $P < 2 \times 10^{-16}$). The F inbreeding coefficient was correlated with the number of homozygous pLoF genes present in each individual. (Spearman r = 0.31; $P = 5 \times 10^{-231}$) (**Fig. 1c**). When restricted to individuals with high levels of inbreeding (F inbreeding coefficient > 6.25%, the expected degree of autozygosity from a first-cousin union), 721 of 1,585 individuals (45%) were homozygous for at least one pLoF mutation.

We tested the hypothesis that genes observed in the homozygous pLoF state in PROMIS participants are under less evolutionary constraint. We calculated the probability of being LoF intolerant (at >90% threshold) for each gene (see Methods) [11,12] and compared this to 1,317 randomly selected genes. The observed 1,317 homozygous pLoF genes were less likely to be classified as highly constrained (odds ratio 0.14; 95% CI 0.12, 0.16; $P < 1 \times 10^{-10}$). Additionally, the 1,317 homozygous pLoF genes are substantially depleted of genes described to be essential for survival and proliferation in four human cancer cell lines (12 of 870 essential genes observed, 1.4%).[13]

A number of genes previously predicted to be required for viability in humans were observed in the homozygous pLoF state in humans (**Supplementary Table 2**). For example, 40 of the 1,317 genes have been associated with embryonic or perinatal lethality as homozygous pLoF in mice.[14] Furthermore, 56 genes predicted to be essential using mouse/human conservation data[15] are tolerated as homozygous pLoF in Pakistani adults. In fact, 9 genes are in both datasets and are also modeled as LoF intolerant.[12] One such gene, *EP400* (also known as *p400*), influences cell cycle regulation via chromatin remodeling[16] and is critical for maintaining the identity of murine embryonic stem cells[17] but we observe an adult human homozygous for disruption of a canonical splice site (intron 3 of 52; c.1435+1G>A) in *EP400*. Conversely, we observed 90 genes where the heterozygous pLoF genotype is of appreciable frequency but the homozygous pLoF genotype is depleted (at P value threshold < 0.05) (**Supplementary Table 3**).

We compared our results to three recent reports where homozygous pLoF genes have been catalogued: in Pakistanis living in Britain, in Icelanders, and in the Exome Aggregation Consortium (ExAC). 3,223 Pakistanis living in Britain with a high degree of parental relatedness (mean 5.62% autozygosity) were sequenced to find 781 homozygous pLoF genes.[18] The sequencing of 2,636 Icelanders and subsequent imputation into 104,220 chip-genotyped Icelanders yielded 1,171 genes in the homozygous pLoF state.[3] Analysis of 52,451 multi-ethnic participants from ExAC (i.e., those not overlapping with current PROMIS study) found 877 genes to be knocked out.[19] Here, we identify a total of 734 unique genes in the homozygous pLoF state that were not observed in the other three studies (**Supplementary Figure 3**).

Intersection of the four sets of genes from these studies revealed only 25 common to all four studies. For example, at phosphodiesteriase 11A (encoded by *PDE11A*), different mutations across the four populations lead to homozygous pLoF state (PROMIS: c.2424-1G>G, p.Cys554ValfsTer14, p.Arg307Ter; ExAC Latino: p.Arg307Ter; ExAC non-Finnish European: p.Cys554ValfsTer14, p.Arg307Ter; Icelanders: p.Arg7ThrfsTer30, p.Arg307Ter; British Pakistani: p.Arg57Ter). The *Pde11a*$^{-}$

$^{/-}$ mouse shows behavioral phenotypes and *PDE11A* is implicated in depression and schizophrenia in humans.[20] Whether humans lacking *PDE11A* also display neuropsychiatric phenotypes remains to be determined.

In order to understand the phenotypic consequences of complete disruption of the 1,317 pLoF genes identified in the PROMIS study, we applied three approaches. First, for 426 genes where two or more participants were homozygous pLoF, we conducted an association screen against a panel of 201 phenotypic traits (**Supplementary Table 4**). Second, in blood samples from each of 84 participants, we measured 1,310 protein biomarkers using a new, multiplexed, aptamer-based proteomics assay. Third, at a single gene, apolipoprotein C-III (encoded by *APOC3*), we recalled participants based on genotype (three classes: 'wild-type', heterozygous pLoF, and homozygous pLoF) and performed provocative physiologic testing.

At 426 genes where two or more participants were homozygous pLoF, we performed association analyses to determine whether homozygous pLoF mutation status was associated with variation in any of 201 traits. For quantitative traits, we compared mean trait values in homozygous pLoF carriers with non-carriers. For dichotomous traits, we performed logistic regression with trait status as the outcome variable and homozygous pLoF carrier status as the predictor variable. Details of covariate adjustments are presented in the Methods. Across quantitative and dichotomous traits, this resulted in the analysis of 18,959 gene-trait pairs and thus, we set Bonferroni-adjusted significance threshold at $P = 3 \times 10^{-6}$.

The quantile-quantile plot of expected versus observed association results shows an excess of highly significant results without systematic inflation (**Supplementary**

Figure 4). Association results surpassed the Bonferroni significance threshold for 26 gene-trait pairs (**Supplementary Table 5**). Below, we highlight seven results: *PLA2G7*, *CYP2F1*, *TREH*, *A3GALT2*, *NRG4*, *SLC9A3R1*, and *APOC3*.

Lipoprotein-associated phospholipase A2 (Lp-PLA2, encoded by *PLA2G7*) hydrolyzes phospholipids to generate lysophosphatidylcholine and oxidized nonesterified fatty acids. In observational epidemiologic studies, higher soluble Lp-PLA2 enzymatic activity has been correlated with increased risk for coronary heart disease; small molecule inhibitors of Lp-PLA2 have been developed for the treatment of coronary heart disease.[21] In PROMIS, we identified participants who are naturally deficient in the Lp-PLA2 enzyme. Two participants are homozygous for a splice-site mutation, *PLA2G7* c.663+1G>A, and 106 are heterozygous for this same mutation. We observed a dose-dependent response relationship between genotype and enzymatic activity: when compared with non-carriers, c.663+1G>A homozygotes have markedly lower Lp-PLA2 enzymatic activity (-245 nmol/ml/min, $P = 2 \times 10^{-7}$) whereas the 106 heterozygotes had an intermediate effect (-120 nmol/ml/min, $P = 2 \times 10^{-77}$) (**Fig. 2a-b**). If Lp-PLA2 plays a causal role for coronary heart disease, one might expect those naturally deficient for this enzyme to have reduced risk for coronary heart disease. We tested the association of *PLA2G7* c.663+1G>A with myocardial infarction across all participants and found that carriers of the pLoF allele did not have reduced risk (OR 0.97; 95% CI, 0.70 – 1.34; P = 0.87) (**Fig. 2c**). In contrast, at two positive control genes, we replicated prior observations (**Supplementary Table 6**); at *LDLR*, heterozygous pLoF mutations increased MI risk 20-fold and, at *PCSK9*, heterozygous pLoF mutations reduced risk by 78%. Of note, in two recent randomized controlled trials, pharmacologic Lp-PLA2 inhibition failed to reduce risk for coronary heart disease,[22,23] a result that might have been anticipated by this genetic analysis.[24]

Cytochrome P450 2F1 (encoded by *CYP2F1*) is primarily expressed in the lung and metabolizes pulmonary-selective toxins, such as cigarette smoke, and thus, modulates the expression of environment-associated pulmonary diseases.[25] At *CYP2F1*, we identified two participants homozygous for a splice-site mutation, c.1295-2A>G. When compared with non-carriers, c.1295-2A>G homozygotes displayed higher soluble interleukin 8 concentrations (3.7-fold increase, $P = 2 \times 10^{-6}$) (**Supplementary Figure 5**). *CYP2F1* c.1295-2A>G heterozygosity had a more modest effect (2.4-fold increase, $P = 2 \times 10^{-4}$). Interleukin 8 induces migration of neutrophils in airways and is a mediator of acute pulmonary inflammation and chronic obstructive pulmonary disease (COPD).[26,27] However, neither carrier reports a personal or family history of obstructive pulmonary disease; further studies of these participants are required to assess the roles of CYP2F1 and interleukin 8 on pulmonary physiology.

Trehalase (encoded by *TREH*) is an intestinal enzyme that splits the naturally-found unabsorbed disaccharide, trehalose, into two glucose molecules.[28] Trehalase deficiency, an autosomal recessive trait, leads to abdominal pain, distention, and flatulence after trehalose ingestion. We identified six participants homozygous for a deletion of a splice acceptor site (c.90-9_106delTCTCTGCAGTGAGATTTACTGCCACG) in exon 2. Homozygotes, unlike heterozygotes or non-carriers, had lower concentrations of several apolipoprotein B-containing lipoprotein subfractions (**Supplementary Table 5**) (**Supplementary Figure 6**).

Alpha-1,3-galactosyltransferase 2 (encoded by *A3GALT2*) catalyzes the formation of the Gal-α1-3Galβ1-4GlcNAc-R (α-gal) epitope; the biological role of this enzyme in humans is uncertain.[29] At *A3GALT2*, we identified two participants homozygous for a frameshift mutation, p.Thr106SerfsTer4. Compared with non-carriers, p.Thr106SerfsTer4 homozygotes both had dramatically reduced concentrations of fasting C-peptide (-97.4%; $P = 6 \times 10^{-12}$) and insulin (-92.3%; $P = 1 \times 10^{-4}$). Such an association was only observed in the homozygous state (**Supplementary Figure 7**). *A3galt2*$^{-/-}$ mice and pigs have recently been shown to have glucose intolerance.[30,31]

To understand if the identification of only a single homozygote may still be informative, we performed a complementary analysis, focusing on those with the most extreme standard Z scores (|Z score| > 5) and requiring that there be evidence for association in heterozygotes as well (see Methods). This procedure highlighted neureglin 4 (*NRG4*), a member of the epidermal growth factor family extracellular ligands which is highly expressed in brown fat, particularly during adipocyte differentiation.[32,33] At *NRG4*, we identified a single participant homozygous for a frameshift mutation, p.Ile75AsnfsTer23, who had nearly absent fasting insulin C-peptide concentrations (-99.3 %; $P = 1 \times 10^{-10}$). When compared with non-carriers, heterozygotes for *NRG4* p.Ile75AsnfsTer23 (n = 8) displayed 48.3 % reduction in insulin C-peptide ($P = 1 \times 10^{-2}$). Mice deleted for *Nrg4* have recently been shown to have glucose intolerance.[33] The single *NRG4* pLoF homozygote participant did not have diabetes nor elevated fasting glucose. Heterozygosity for a *NRG4* pLoF mutation (n=26) was also not associated with diabetes or fasting glucose. More detailed phenotyping will be required to definitively assess any relationship of *NRG4* deficiency in humans with glucose intolerance.

To further dissect the consequences of a subset of homozygous pLoF genes, we measured 1,310 protein biomarkers in 84 participants through a new, multiplexed, proteomic assay (SOMAscan). Among the 84 participants, there were nine genes with at least two pLoF homozygotes and we associated these genotypes across 1,310 protein biomarkers and observed a number of associations (**Supplementary Table 7**). We highlight two PROMIS participants who are homozygous pLoF at *SLC9A3R1*; these participants have increased circulating concentrations of several proteins involved in parathyroid hormone or osteoclast signaling including calcium / calmodulin-dependent protein kinase II (CAMK2) alpha, beta, and delta subunits, cAMP-regulated phosphoprotein 19, and signal transducer and activator of transcription (STAT) 1, 3, and 6 (**Supplementary Table 7**). *SLC9A3R1* (aka *NHERF1*) encodes a Na+/H+ exchanger regulatory cofactor that interacts with and regulates the parathyroid hormone receptor; *Nherf1*$^{-/-}$ mice display hyperphosphaturia and disrupted protein kinase A-dependent cAMP-mediated phosphorylation.[34,35] Humans carrying rare missense mutations in *SLC9A3R1* have nephrolithiasis, osteoporosis, and hypophosphatemia.[36]

Apolipoprotein C-III (apoC-III, encoded by *APOC3*) is a major protein component of chylomicrons, very low-density lipoprotein cholesterol, and high-density lipoprotein cholesterol.[37] We and others recently reported that *APOC3* pLoF mutations in heterozygous form lower plasma triglycerides and reduce risk for coronary heart disease[5,6,38]; there is now substantial interest in *APOC3* as a therapeutic target.[39-41] In published studies, no *APOC3* pLoF homozygotes have been identified despite study of nearly 200,000 participants from the U.S. and Europe, raising concerns that complete *APOC3* deficiency may be harmful. However, in our study of ~10,000 Pakistanis, we identified four participants homozygous for *APOC3* p.Arg19Ter. When compared with non-carriers, p.Arg19Ter homozygotes displayed near-absent plasma apoC-III protein (-88.9 %, $P = 5 \times 10^{-23}$), lower plasma triglyceride concentrations (-59.6 %, $P = 7 \times 10^{-4}$), higher high-density lipoprotein (HDL) cholesterol (+26.9 mg/dL, $P = 3 \times 10^{-8}$); and similar levels of low-density lipoprotein (LDL) cholesterol ($P = 0.14$) (**Fig. 3a-d**).

ApoC-III functions as a brake on the metabolism of dietary fat and thus, the complete lack of this protein should promote handling of ingested fat. We re-contacted one homozygous pLoF proband, his wife, and 27 of his first-degree relatives for genotyping and physiologic investigation. We found that the proband's wife, a first cousin, was also a pLoF homozygote, leading to all nine children being obligate homozygotes (**Fig. 3e**). In this family, we challenged pLoF homozygotes (*APOC3*$^{-/-}$; n = 6) and non-carriers (*APOC3*$^{+/+}$; n = 7) with a 50 g/m^2 oral fat load followed by serial blood testing for six hours. *APOC3* p.Arg19Ter homozygotes had significantly lower post-prandial triglyceride excursions (triglycerides area under the curve 468.3 mg/dL*6 hours vs 1267.7 mg/dL*6 hours; $P = 1 \times 10^{-4}$) (**Fig. 3f**). These data show that complete lack of apoC-III markedly improves clearance of plasma triglycerides after a fatty meal and are consistent with and extend an earlier report of diminished post-prandial lipemia in *APOC3* pLoF heterozygotes.[38]

Targeted gene disruption in model organisms followed by phenotypic analysis has been a fruitful approach to understand gene function[42]; here, we extend this concept to the human organism, leveraging naturally-occurring pLoF mutations, consanguinity, and biochemical phenotyping. These results permit several conclusions. First, power to identify human knockouts is improved with the study of multiple populations and particularly those with high degrees of consanguinity. Using the observed median inbreeding coefficient of sequenced participants and genotypes from the first 7,078 sequenced Pakistanis, we estimate that the sequencing of 200,000 Pakistanis, may result in up to 8,754 genes (95% CI, 8,669-8,834) completely knocked out in at least one participant (**Fig. 4**).

Second, a panel of phenotypes measured in a blood sample can yield hypotheses regarding phenotypic consequences of gene disruption as observed for *PLA2G7*, *CYP2F1*, *TREH*, *A3GALT2, NRG4*, *SLC9A3R1*, and *APOC3*. Finally, recall by genotype followed by provocative testing may provide physiologic insights. We used this approach to demonstrate that complete lack of apoC-III is tolerated and results in both lowered fasting triglyceride concentrations as well as substantially blunted post-prandial lipemia.

Several limitations deserve mention. First and most importantly, any given mutation annotated as pLoF may not truly lead to loss of protein function. In addition to bioinformatics filtration, we manually curated all homozygous pLoF variants (n=1,580) to assess confidence in variant fidelity and predicted biochemical impact (**Supplementary Table 1** and **Supplementary Table 8**). We found 56 variants with genotypes with a low number of supportive reads, 55 with poorly mapped reads (**Supplementary Table 9**), and an additional 66 where there were potential mechanisms of protein-truncation rescue (**Supplementary Figure 8**) or occurred within exons or splice sites where conservation was low. Thus, we found the majority of pLoF calls (1403 out of 1580; 89%) to be free of mapping or annotation error. However, for any given pLoF, experimental validation will be required to prove loss of gene function (e.g., targeted assays such as RT-PCR of transcript and/or Western blot of protein to confirm its absence). Second, statistical power for genotype-phenotype correlation is low if a gene is knocked-out in only 1 or 2 participants. However, this situation should improve with larger sample sizes (**Supplementary Figure 9**). Third, statistical power in the proteomics analysis may be low because of the limited number of samples assayed and the impact of non-genetic factors on plasma concentrations.[43] Finally, our analysis was limited to available phenotypes and in only one instance did we recall participants for deeper phenotyping; rather, a standardized clinical phenotyping protocol is desirable for each participant where a gene is observed to be knocked out.

To date, most human genetic studies have pursued a phenotype-first ("forward" genetics) approach, beginning with traits of interest followed by genetic mapping. It is now feasible to pursue a systematic genotype-first ("reverse" genetics) approach, starting with homozygous pLoF humans followed by methodical examination of a diverse set of traits.

These observations set the stage for a 'human knockout project,' a systematic effort to understand the phenotypic consequences of complete disruption of every gene in the human genome. Key elements for a human knockout project include: 1) identification of populations where homozygous genotypes may be enriched[18,44]; 2) deep-coverage sequencing of the protein-coding regions of the genome[3]; 3) availability of a broad array biochemical as well as clinical phenotypes across the population; 4) ability to re-contact knockout humans as well as family members; 5) a thorough clinical evaluation in each participant where a gene is observed to be knocked out; and 6) hypothesis-driven provocative phenotyping in selected participants.

Methods

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

The PROMIS study was designed to investigate determinants of cardiometabolic diseases in Pakistan. Since 2005, the study has enrolled close to 38,000 participants; the present investigation sequenced 10,503 participants selected as 4,793 cases with myocardial infarction and 5,710 controls free of myocardial infarction. Participants aged 30-80 years were enrolled from nine recruitment centers based in five major urban cities in Pakistan. Type 2 diabetes in the study was defined based on self-report or fasting glucose levels >125 mg/dL or HbA1c > 6.5 % or use of glucose lowering medications. The institutional review board at the Center for Non-Communicable Diseases (IRB: 00007048, IORG0005843, FWAS00014490) approved the study and all participants gave informed consent.

Phenotype descriptions.

Non-fasting blood samples (with the time since last meal recorded) were drawn and centrifuged within 45 minutes of venipuncture. Serum, plasma and whole blood samples were stored at -70°C within 45 minutes of venipuncture. All samples were transported on dry ice to the central laboratory at the Center for Non-Communicable Diseases (CNCD), Pakistan, where serum and plasma samples were aliquoted across 10 different storage vials. Samples were stored at -70°C for any subsequent laboratory analyses. All biochemical assays were conducted in automated auto-analyzers. At CNCD Pakistan, measurements for total-cholesterol, HDL cholesterol, LDL cholesterol, triglycerides, and creatinine were made in serum samples using enzymatic assays; whereas levels of HbA1c were measured using a turbidemetric assay in whole-blood samples (Roche Diagnostics, USA). For further measurements, aliquots of serum and plasma samples were transported on dry ice to the Smilow Research Center, University of Pennsylvania, USA, where following biochemical assays were conducted: apolipoproteins (apoA-I, apoA-II, apoB, apoC-III, apoE) and non-esterified fatty acids were measured through immunoturbidometric assays using kits by Roche Diagnostics or Kamiya; lipoprotein (a) levels were determined through a turbidimetric assay using reagents and calibrators from Denka Seiken (Niigata, Japan); LpPLA2 mass and activity levels were determined using immunoassays manufactured by diaDexus (San Francisco, CA, USA); measurements for insulin, leptin and adiponectin were made using radio-immunoassays by LINCO (MO, USA); levels of adhesion molecules (ICAM-1, VCAM-1, P- and E-Selectin) were determined through enzymatic assays by R&D (Minneapolis, MN, USA); and measurements for C-reactive protein, alanine transaminase, aspartate transaminase, cystatin-C, ferritin, ceruloplasmin, thyroid stimulating hormone, alkaline phosphatase, sodium, potassium, choloride, phosphate, sex-harmone binding globulin were made using enzymatic assays manufactured by Abbott Diagnostics (NJ, USA). Glomerular filtration rate (eGFR) was estimated from serum creatinine levels using the MDRD equation. ApoC-III levels were determined in an autoanalyzer using a commercially available ELISA by Sekisui Diagnostics (Lexington, USA). We also measured the following 52 protein biomarkers by multiplex immunoassay using a customised panel on the Luminex 100/200 instrument by RBM (Myriad Rules Based Medicine, Austin, TX, USA): fatty acid binding protein, granuloctye monocyte colony stimulating factor, granulocyte colony stimulating factor, interferon gamma, interleukin-1 beta, interleukin 1 receptor, interleukin 2, interleukin 3, interleukin 4, interleukin 5, interleukin 6, interleukin 7, interleukin 8, interleukin 10, interleukin 18, interleukin p40, interleukin p70, interleukin 15, interleukin 17, interleukin 23, macrophage inflammatory protein 1 alpha, macrophage inflammatory protein 1 beta, malondialdehyde-modified LDL, matrix metalloproteinase 2, matrix metalloproteinase 3, matrix metalloproteinase 9, nerve growth factor beta, tumor necrosis factor alpha, tumor necrosis factor beta, brain-derived neurotrophic factor, CD40, CD40 ligand, eotaxin, factor VII, insulin-like growth factor 1, lecithin-type oxidized LDL receptor 1, monocyte chemoattractant protein 1, myeloperoxidase, N-terminal prohormone of brain natriuretic peptide, neuronal cell adhesion molecule, pregnancy-associated plasma protein A, soluble receptor for advanced glycation end-products, sortilin, stem cell factor, stromal cell-derived factor 1, thrombomodulin, S100 calcium binding protein B, and vascular endothelial growth factor.

Laboratory methods for array-based genotyping.

As previously described, a genomewide association scan was performed using the Illumina 660 Quad array at the Wellcome Trust Sanger Institute (Hinxton, UK) and using the Illumina HumanOmniExpress at Cambridge Genome Services, UK.[45] Initial quality control (QC) criteria included removal of participants or single nucleotide polymorphisms (SNPs) that had a missing rate >5%. SNPs with a MAF <1% and a P-value of $<10^{-7}$ for the Hardy-Weinberg equilibrium test were also excluded from the analyses. In PROMIS, further QC included removal of participants with discrepancy between their reported sex and genetic sex determined from the X chromosome. To identify sample duplications, unintentional use of related samples (cryptic relatedness) and sample contamination (individuals who seem to be related to nearly everyone in the sample), identity-by-descent (IBD) analyses were conducted in PLINK.[46]

Laboratory methods for exome sequencing.

Exome sequencing. Exome sequencing was performed at the Broad Institute. Sequencing and exome capture methods have been previously described.[47,48] A brief description of the methods is provided below.

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

Library construction. Library construction was performed as previously described[49], with the following modifications: initial genomic DNA input into shearing was reduced from 3µg to 10-100ng in 50µL of solution. For adapter ligation, Illumina paired end adapters were replaced with palindromic forked adapters, purchased from Integrated DNA Technologies, with unique 8 base molecular barcode sequences included in the adapter sequence to facilitate downstream pooling. With the exception of the palindromic forked adapters, the reagents used for end repair, A-base addition, adapter ligation, and library enrichment PCR were purchased from KAPA Biosciences (Wilmington, MA, USA) in 96-reaction kits. In addition, during the post-enrichment SPRI cleanup, elution volume was reduced to 20 μL to maximize library concentration, and a vortexing step was added to maximize the amount of template eluted.

In-solution hybrid selection. 1,970 samples underwent in-solution hybrid selection as previously described[49], with the following exception: prior to hybridization, two normalized libraries were pooled together, yielding the same total volume and concentration specified in the publication. 8,808 samples underwent hybridization and capture using the relevant components of Illumina's Rapid Capture Exome Kit and following the manufacturer's suggested protocol, with the following exceptions: first, all libraries within a library construction plate were pooled prior to hybridization, and second, the Midi plate from Illumina's Rapid Capture Exome Kit was replaced with a skirted PCR plate to facilitate automation. All hybridization and capture steps were automated on the Agilent Bravo liquid handling system.

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

Cluster amplification and sequencing. Cluster amplification of denatured templates was performed according to the manufacturer's protocol (Illumina) using HiSeq v3 cluster chemistry and HiSeq 2000 or 2500 flowcells. Flowcells were sequenced on HiSeq 2000 or 2500 using v3 Sequencing-by-Synthesis chemistry, then analyzed using RTA v.1.12.4.2. Each pool of whole exome libraries was run on paired 76bp runs, with and 8 base index sequencing read was performed to read molecular indices, across the number of lanes needed to meet coverage for all libraries in the pool.

Read mapping and variant discovery. Samples were processed from real-time base-calls (RTA v.1.12.4.2 software [Bustard], converted to qseq.txt files, and aligned to a human reference (hg19) using Burrows–Wheeler Aligner (BWA).[50] Aligned reads duplicating the start position of another read were flagged as duplicates and not analysed. Data was processed using the Genome Analysis ToolKit (GATK v3).[51-53] Reads were locally realigned around insertions-deletions (indels) and their base qualities were recalibrated. Variant calling was performed on both exomes and flanking 50 base pairs of intronic sequence across all samples using the HaplotypeCaller (HC) tool from the GATK to generate a gVCF. Joint genotyping was subsequently performed and 'raw' variant data for each sample was formatted (variant call format (VCF)). Single nucleotide polymorphisms (SNVs) and indel sites were initially filtered after variant calibration marked sites of low quality that were likely false positives.

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

Data processing and quality control of exome sequencing.

Variant annotation. Variants were annotated using Variant Effect Predictor[54] and the LOFTEE[10] plugin to identify protein-truncating variants predicted to disrupt the respective gene's function with "high confidence." Each allele at polyallelic sites was separately annotated.

Sample level quality control. We performed quality control of samples using the following steps. For quality control of samples, we used bi-allelic SNVs that passed the GATK VQSR filter and were on genomic regions targeted by both ICE and Agilent exome captures. We removed samples with discordance rate > 10% between genotypes from exome sequencing with genotypes from array-based genotyping and samples with sex mismatch between inbreeding coefficient on chromosome X and fingerprinting. We tested for sample contamination using the verifyBamID software, which examines the proportion of non-reference bases at reference sites, and excluded samples with high estimated contamination (FREEMIX scores > 0.2).[55] After removing monozygotic twins or duplicate samples using the KING software[56], we removed outlier samples with too many or too few SNVs (>17,000 or <12,000 total variants; >400 singletons; and >300 doubletons). We removed those with extreme overall transition-to-transversion ratios (>3.8 or <3.3) and heterozygosity (heterozygote:non-reference homozygote ratio >6 or <2). Finally, we removed samples with high missingness (>0.05).

Variant level quality control. Variant score quality recalibration was performed separately for SNVs and indels use the GATK VariantRecalibrator and ApplyRecalibration to filter out variants with lower accuracy scores. Additionally, we removed sites with an excess of heterozygosity calls (InbreedingCoeff <-0.3). To further reduce the rate of inaccurate variant calls, we further filtered out SNVs with low average quality (quality per depth of coverage (QD) < 2) and a high degree of missingness (> 20 %), and indels also with low average quality (quality per depth of coverage (QD) < 3) and a high degree of missingness (> 20 %).

Laboratory methods for proteomics.

Protein capture. For 91 participants, enriched for homozygous pLoF mutations, we measured 1,310 protein analytes in plasma using the SOMAscan assay (SomaLogic, Boulder, CO, USA). Protein-capture was performed using modified aptamer technology as previously described.[57] Briefly, modified nucleotides, analogous to antibodies, on a custom DNA microarray recognize intact tertiary protein structures. After washing, complexes are released from beads by photocleavage of the linker with UV light and the resultant relative fluorescent unit is proportional to target protein.

Quality control. Samples (n = 7) were excluded if they showed evidence of systematic inflation of association, or >5 % of traits in the top or bottom 1st percentile of the analytic distribution.

Methods for manual curation of a subset of pLoF **variants.**

Manual curation was performed collaboratively by three geneticists: 25 pLoF variant calls were reviewed independently by two reviewers and compared to ensure similar review criteria before the remainder was divided and separately assessed by each of the two reviewers separately. A third reviewer resolved discrepancies. Read and genotype support was confirmed by review of reads in Integrative Genomics Viewer. We flagged pLoF variants for any of the following six reasons: 1) read-mapping flags; 2) genotyping flags; 3) presence of an additional polymorphism which rescues protein truncation; 4) presence of an additional polymorphism which rescues splice site; 5) if affecting a minority of transcripts; and 6) polymorphism occurs at exon or splice site with low conservation. Criteria for these reasons are provided in **Supplementary Table 8**.

Methods for inbreeding analyses.

Array-derived runs of homozygosity. Analyses were conducted in PLINK[46] using genome-wide association (GWAS) data in PROMIS and HapMap 3 populations. Segments of the genome that were at-least 1.5 Mb long, had a SNP density of 1 SNP per 20 kb and had 25 consecutive homozygous SNPs (1 heterozygous and/or 5 missing SNPs were permitted within a segment) were defined to be in a homozygous state (or referred as "runs of homozygosity" (ROH)), as described previously.[58] Homozygosity was expressed as the percentage of the autosomal genome found in a homozygous state, and was calculated by dividing the sum of ROH length within each individual by the total length of the autosome in PROMIS and HapMap 3 populations respectively. To investigate variability in homozygosity explained by parental consanguinity, the difference in R^2 is reported for a linear regression model of homozygosity including and excluding parental consanguinity on top of age, sex and the first 10 principal components derived from the typed autosomal GWAS data.

Sequencing-derived coefficient of inbreeding. We compared the coefficient of inbreeding distributions of 10,503 exome sequenced PROMIS participants with 15,248 participants (European ancestry = 12,849, and African ancestry = 2,399) who were exome sequenced at the Broad Institute (Cambridge, MA) from the Myocardial Infarction Genetics consortium.[48] We extracted approximately 5,000 high-quality polymorphic SNVs in linkage equilibrium present on both target intervals that passed variant quality control metrics based on HapMap 3 data.[59] Using PLINK, we estimated the coefficient of inbreeding separately within each ethnicity group.[46] The coefficient of inbreeding was estimated as the observed degree of homozygosity compared with the anticipated homozygosity derived from an estimated common ancestor.[60] The Wilcoxon-Mann-Whitney test was used to test whether PROMIS participants had different median coefficients of inbreeding compared to other similarly sequenced outbred individuals and whether the median coefficient of inbreeding was different between PROMIS participants who reported parental relatedness versus not. A two-sided *P* of 0.05 was the pre-specified threshold for statistical significance.

Methods for sequencing projection analysis.

To compare the burden of unique completely inactivated genes in the PROMIS cohort with outbred cohorts of diverse ethnicities, we extracted the minor allele frequencies (maf) of "high confidence" loss-of-function mutations observed in the first 7,078 sequenced PROMIS participants, and in European, African, and East Asian ancestry participants from the Exome Aggregation Consortium (ExAC r0.3; exac.broadinstitute.org). For each gene and for each ethnicity, the combined minor allele frequency (cmaf) of rare (maf < 0.1%) “high confidence” loss-of-function mutations was calculated. We then simulated the number of unique completely inactivated genes across a range of sample sizes per ethnicity and PROMIS. The expected probability of observing complete inactivation (two pLoF copies in an individual) of a gene was calculated as $(1 - F) * cmaf^{\ 2} + \ F * cmaf$, which accounts for allozygous and autozygous, respectively, mechanisms for complete genie knockout. F, the inbreeding coefficient, is defined as $F = 1 - (expected\ heterozygosity\ rate\ /\ observed\ heterozygosity\ rate)$. For PROMIS, the median F inbreeding coefficient (0.016) was used for estimation. Down-sampling within the observed sample size for both high-confidence pLoF mutations and synonymous variants did not deviate significantly from the expected trajectory (**Supplementary Figure 11**). For a range of sample sizes (0-200,000), each gene was randomly sampled under a binomial distribution ($X \sim B(n, cmaf)$) and it was determined if the gene was successfully sampled at least once. To refine the estimated count of unique genes per sample size, each sampling was replicated ten times.

Methods for constraint score analysis.

We sought to determine whether the observed homozygous pLoF genes were under less evolutionary constraint by first obtaining constraint loss of function constraint scores derived from the Exome Aggregation Consortium (Lek M et al, in preparation).[11,12] Briefly, we used the number of observed and expected rare (MAF < 0.1%) loss of function variants per gene to determine to which of three classes it was likely to belong: pLoF (observed variation matches expectation), recessive (observed variation is ~50% expectation), or haploinsufficient (observed variation is <10% of expectation). The probability of being loss of function intolerant (pLI) of each transcript was defined as the probability of that transcript falling into the haploinsufficient category. Transcripts with a pLI ≥ 0.9 are considered very likely to be loss of function intolerant; those with pLI ≤ 0.1 are not likely to be loss of function intolerant. A list of 1,317 genes were randomly sampled from a list of sequenced genes 1,000 times and the proportion of loss of function intolerant genes compared to the proportion of the observed homozygous pLoF genes was compared using the chi square test. The likelihood that the distribution of the test statistics deviated from the pLoF was ascertained.

Additionally, we sought to determine whether there were genes with appreciate pLoF allele frequencies yet relative depletion of homozygous pLoF genotypes. We computed estimated genotype frequencies based on Hardy-Weinberg equilibrium and the F inbreeding coefficient and compared the frequencies to the observed genotype counts with the chi square goodness-of-fit test. A nominal $P < 0.05$ is used to demonstrate at least nominal association.

Methods for rare variant association analysis.

Recessive model association discovery. We sought to determine whether complete loss-of-function of a gene was associated with a dense array of phenotypes. We extracted a list of individuals per gene who were homozygous for a high confidence pLoF allele that was rare (minor allele frequency < 1 %) in the cohort. From a list of 1,317 genes where there was at least one participant homozygous pLoF and a list of 201 traits, we initially considered 264,717 gene-trait pairings. To reduce the likelihood of false positives, we only considered gene-trait pairs where there were at least two homozygous pLoF alleles per gene phenotyped for a given trait yielding 18,959 gene-trait pairs for analysis. For all analyses, we constructed generalized linear models to test whether complete loss of function versus non-carriers was associated with trait variation. A logit link was used for binomial outcomes. Right-skewed continuous traits were natural log transformed. Age, sex, and myocardial infarction status were used as covariates in all analyses. We extracted principal components of ancestry using EIGENSTRAT to control for population stratification in all analyses.[61] For lipoprotein-related traits, the use of lipid-lowering therapy was used as a covariate. For glycemic biomarkers, only non-diabetics were used in the analysis. The P threshold for statistical significance was 0.05 / 18,959 = 3 x 10^{-6}.

Heterozygote association replication. We hypothesized that some of the associations for homozygous pLoF alleles will display a more modest effect for heterozygous pLoF alleles. Thus, the aforementioned analyses were performed comparing heterozygous pLoF carriers to non-carriers for the 26 homozygous pLoF-trait associations that surpassed prespecified statistical significance. A P of 0.05 / 26 = 0.002 was set for statistical significance for these restricted analyses.

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

Recessive model association discovery for proteomics. Among the 84 participants with proteomic analyses of 1,310 protein analytes, 9 genes were observed in the homozygous pLoF state at least twice. We log transformed each analyte and associated with homozygous pLoF genotype status, adjusting for proteomic plate, age, sex, myocardial infarction status, and principal components. Gene-analyte associations were considered significant if P values were less than 0.05 / (1,310 x 9) = 4.3 x 10^{-6}.

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

Methods for Sanger sequencing. We collected blood samples from a total of 28 subjects, including one of the four *APOC3* p.Arg19Ter homozygous participants along with 27 of his family and community members for DNA extraction and separated into plasma for lipid and apolipoprotein measurements. All subjects were consented prior to initiation of the studies (IRB: 00007048 at the Center for Non-Communicable Diseases, Paksitan). DNA was isolated from whole blood using a reference phenol-chloroform protocol.[62] Genotypes for the p.Arg19Ter variant were determined in all 28 participants by Sanger sequencing. A 685 bp region of the *APOC3* gene including the base position for this variant was amplified by PCR (Expand HF PCR Kit, Roche) using the following primer sequences: Forward primer CTCCTTCTGGCAGACCCAGCTAAGG, Reverse primer CCTAGGACTGCTCCGGGGAGAAAG. PCR products were purified with Exo-SAP-IT (Affymetrix) and sequenced via Sanger sequencing using the same primers.

Oral fat tolerance test. Six non-carriers and seven homozygotes also participated in an oral fat tolerance test. Participants fasted overnight and then blood was drawn for measurement of baseline fasted lipids. Following this, participants were administered an oral load of heavy cream (50 g fat per square meter of body surface area as calculated by the method of Mosteller[63]). Participants consumed this oral load within a time span of 20 minutes and afterwards consumed 200 mL of water. Blood was drawn at 2, 4, and 6 hours after oral fat consumption as done previously.[38,64] All lipid and apolipoprotein measurements from these plasma samples were determined by immunoturbidimetric assays on an ACE Axcel Chemistry analyzer (Alfa Wasserman). A comparisons of area-under-the curve triglycerides was performed between *APOC3* p.Arg19Ter homozygotes and non-carriers using a two independent sample Student's t test; $P < 0.05$ was considered statistically significant.

Tables

Table 1. Baseline characteristics of exome sequenced study participants.

Characteristic	**Value**
	(n = 10,503)
Age (yrs) – mean (sd)	52.0 (9.0)
Women – no. (%)	1,802 (17.2 %)
Parents closely related – no. (%)	4,101 (39.0 %)
Spouse closely related – no. (%)	4,182 (39.8 %)
Ethnicity – no. (%)	
Urdu	3,846 (36.6 %)
Punjabi	3,668 (34.9 %)
Sindhi	1,128 (10.7 %)
Pathan	589 (5.6 %)
Memon	141 (1.3 %)
Gujarati	109 (1.0 %)
Balochi	123 (1.2 %)
Other	891 (8.5 %)
Hypertension – no. (%)[*]	4,744 (45.2 %)
Hypercholesterolemia – no. (%)[†]	2,924 (27.8 %)
Diabetes mellitus – no. (%)[‡]	4,264 (40.6 %)
Coronary heart disease – no. (%)[§]	4,793 (45.6 %)
Smoking – no. (%)[‖]	4,201 (40.0 %)

BMI (m/kg^2) – mean (sd)	25.9 (4.2)

[*] Hypertension defined as systolic blood pressure ≥140 mmHg, diastolic blood pressure ≥90 mmHg, or antihypertensive treatment.

[†] Hypercholesterolemia defined as serum total cholesterol >240 mg/dL, lipid lowering therapy or self-report.

[‡] Diabetes defined as fasting blood glucose ≥126 mg/dL, or HbA1c >6.5 %, oral hypoglycemics, insulin treatment, or self-report.

[§] Coronary heart disease defined as acute myocardial infarction as determined by clinical symptoms with typical EKG findings or elevated serum troponin I.

[||] Smoking defined as active current or prior tobacco smoking.

<u>Figure Legends</u>

Fig 1. a, Most genes are observed in the homozygous pLoF state in only single individuals. **b.** The distribution of F inbreeding coefficient of PROMIS participants is compared to those of outbred samples of African (AFR) and European (EUR) ancestry. **c**, The burden of homozygous pLoF genes per individual is correlated with coefficient of inbreeding.

Fig 2. a.-b. Carriage of a splice-site mutation, c.663+1G>A, in *PLA2G7* leads to a dose-dependent reduction of both lipoprotein-associated phospholipase A2 (Lp-PLA2) mass and activity, with homozygotes having no circulating Lp-PLA2. **c.** Despite substantial reductions of Lp-PLA2 activity, *PLA2G7* c.663+1G>A heterozygotes and homozygotes have similar coronary heart disease risk when compared with non-carriers.

Fig 3. a.-d. *APOC3* pLoF genotype status, apolipoprotein C-III, triglycerides, HDL cholesterol and LDL cholesterol distributions among all sequenced participants. Apolipoprotein C-III concentration is displayed on a logarithmic base 10 scale. **e.** A proband with *APOC3* pLoF homozygote genotype as well as several family members were recalled for provocative phenotyping. Surprisingly, the spouse of the proband was also a pLoF homozygote, leading to nine obligate homozygote children. Given the extensive first-degree unions, the pedigree is simplified for clarity. **f.** *APOC3* p.Arg19Ter homozygotes and non-carriers within the same family were challenged with a 50 g/m^2 fat feeding. Homozygotes had lower baseline triglyceride concentrations and displayed marked blunting of post-prandial rise in plasma triglycerides.

Fig 4. Number of unique homozygous pLoF genes anticipated with increasing sample sizes sequenced in PROMIS compared with similar African (AFR) and European (EUR) sample sizes. Estimates derived using observed allele frequencies and degree of inbreeding.

Figures

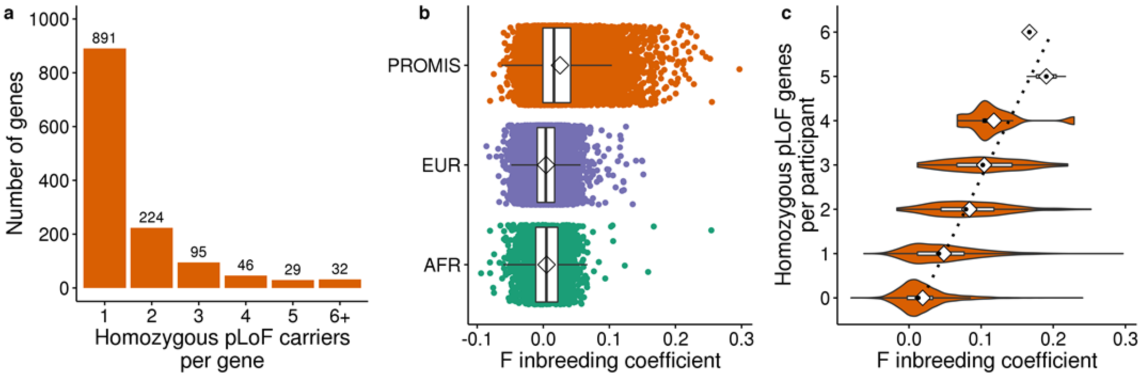

Fig. 1. Homozygous pLoF burden in PROMIS is driven by excess autozygosity.

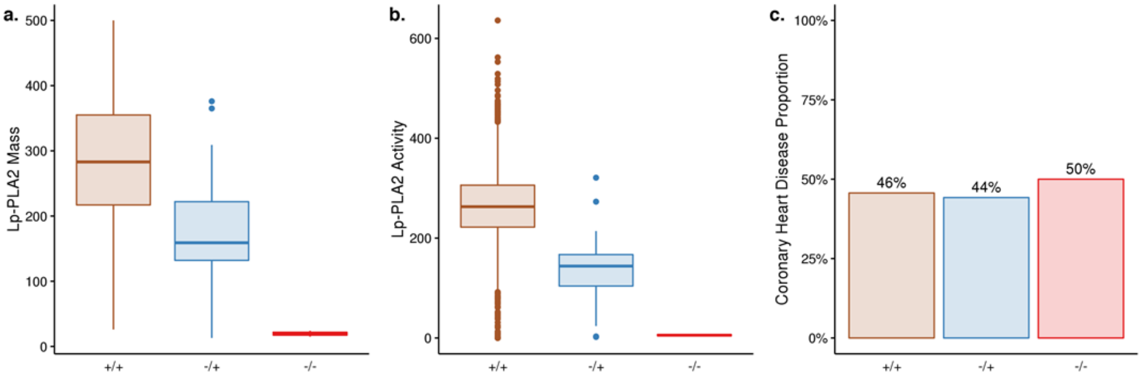

Fig. 2. Carriers of *PLA2G7* splice mutation have diminished Lp-PLA2 mass (*P* = 6 x 10^{-5}) and activity (*P* = 2 x 10^{-7}) but similar risk for coronary heart disease risk when compared to non-carriers (*P* = 0.87).

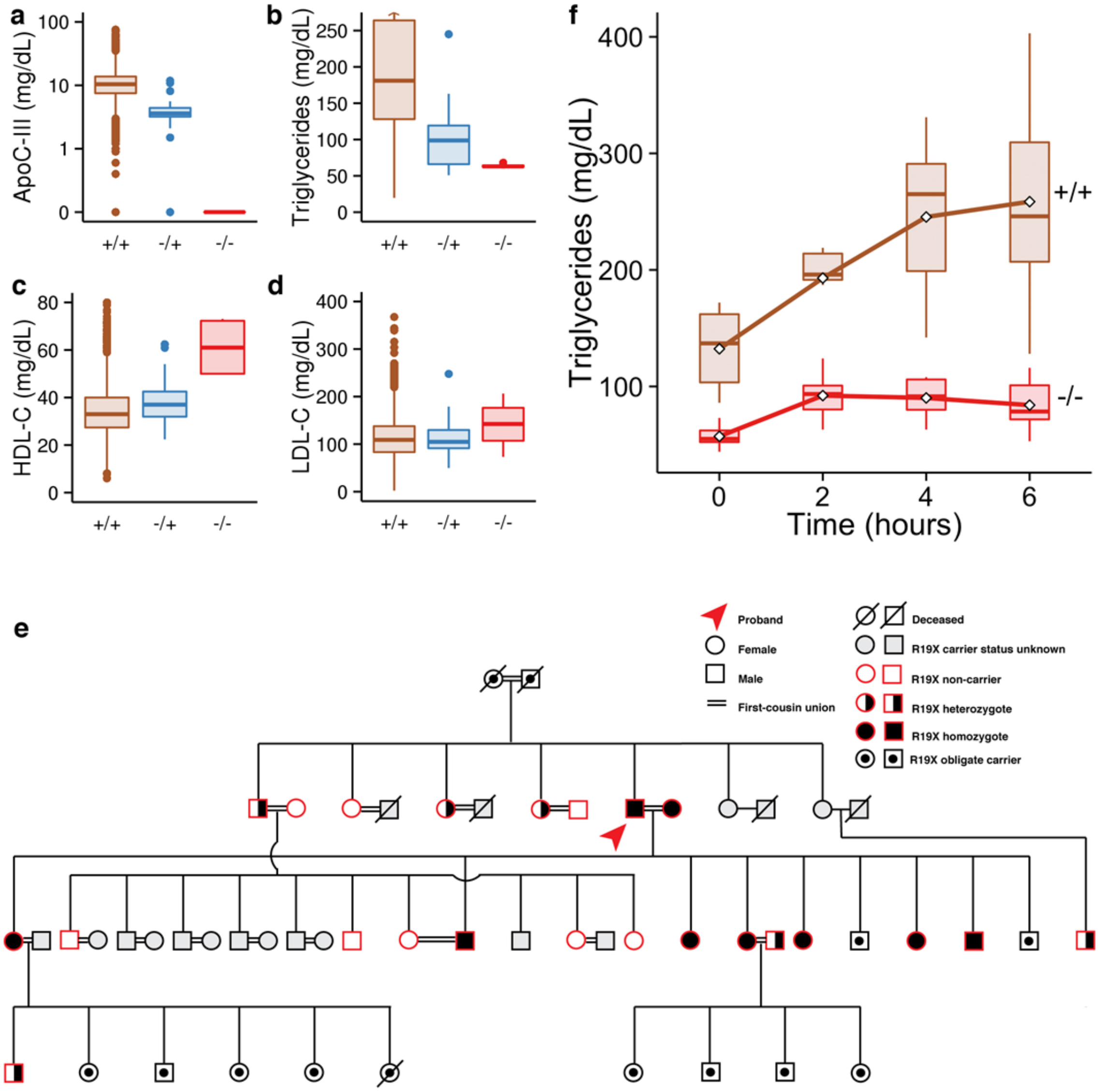

Fig 3. *APOC3* pLoF homozygotes have diminished fasting triglycerides and blunted post-prandial lipemia.

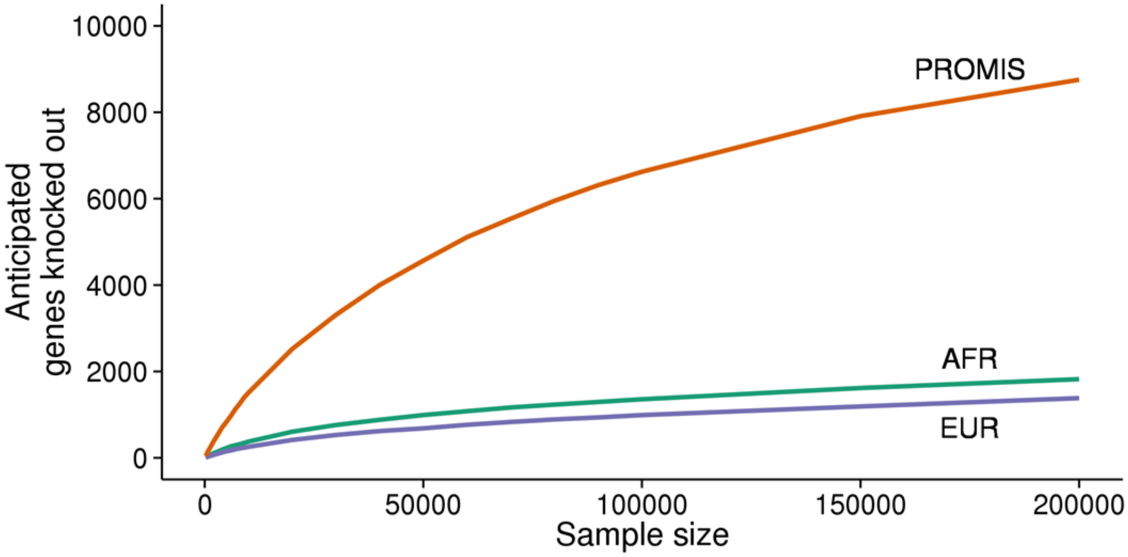

Fig 4. Simulations anticipate many more homozygous pLoF genes in the PROMIS cohort.

References

1 Eisenberg, D., Marcotte, E. M., Xenarios, I. & Yeates, T. O. Protein Function in the Post-Genomic Era. *Nature* **405**, 823-826, doi:10.1038/35015694 (2000).

2 MacArthur, D. G. *et al.* A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. *Science* **335**, 823-828, doi:10.1126/science.1215040 (2012).

3 Sulem, P. *et al.* Identification of a Large Set of Rare Complete Human Knockouts. *Nat Genet*, doi:10.1038/ng.3243 (2015).

4 Bittles, A. H., Mason, W. M., Greene, J. & Rao, N. A. Reproductive Behavior and Health in Consanguineous Marriages. *Science* **252**, 789-794 (1991).

5 Crosby, J. *et al.* Loss-of-Function Mutations in Apoc3, Triglycerides, and Coronary Disease. *N Engl J Med* **371**, 22-31, doi:10.1056/NEJMoa1307095 (2014).

6 Jorgensen, A. B., Frikke-Schmidt, R., Nordestgaard, B. G. & Tybjaerg-Hansen, A. Loss-of-Function Mutations in Apoc3 and Risk of Ischemic Vascular Disease. *N Engl J Med* **371**, 32-41, doi:10.1056/NEJMoa1308027 (2014).

7 Saleheen, D. *et al.* The Pakistan Risk of Myocardial Infarction Study: A Resource for the Study of Genetic, Lifestyle and Other Determinants of Myocardial Infarction in South Asia. *Eur J Epidemiol* **24**, 329-338, doi:10.1007/s10654-009-9334-y (2009).

8 Modell, B. & Darr, A. Science and Society: Genetic Counselling and Customary Consanguineous Marriage. *Nat Rev Genet* **3**, 225-229, doi:10.1038/nrg754 (2002).

9 Lander, E. S. & Botstein, D. Homozygosity Mapping: A Way to Map Human Recessive Traits with the DNA of Inbred Children. *Science* **236**, 1567-1570 (1987).

10 Karczewski, K. J. *Loftee (Loss-of-Function Transcript Effect Estimator)*, <https://github.com/konradjk/loftee> (2015).

11 De Rubeis, S. *et al.* Synaptic, Transcriptional and Chromatin Genes Disrupted in Autism. *Nature* **515**, 209-215, doi:10.1038/nature13772 (2014).

12 Samocha, K. E. *et al.* A Framework for the Interpretation of De Novo Mutation in Human Disease. *Nat Genet* **46**, 944-950, doi:10.1038/ng.3050 (2014).

13 Wang, T. *et al.* Identification and Characterization of Essential Genes in the Human Genome. *Science* **350**, 1096-1101, doi:10.1126/science.aac7041 (2015).

14 Eppig, J. T. *et al.* The Mouse Genome Database (Mgd): Facilitating Mouse as a Model for Human Biology and Disease. *Nucleic Acids Res* **43**, D726-736, doi:10.1093/nar/gku967 (2015).

15 Georgi, B., Voight, B. F. & Bucan, M. From Mouse to Human: Evolutionary Genomics Analysis of Human Orthologs of Essential Genes. *PLoS Genet* **9**, e1003484, doi:10.1371/journal.pgen.1003484 (2013).

16 Fuchs, M. *et al.* The P400 Complex Is an Essential E1a Transformation Target. *Cell* **106**, 297-307 (2001).

17 Fazzio, T. G., Huff, J. T. & Panning, B. An Rnai Screen of Chromatin Proteins Identifies Tip60-P400 as a Regulator of Embryonic Stem Cell Identity. *Cell* **134**, 162-174, doi:10.1016/j.cell.2008.05.031 (2008).

18 Narasimhan, V. M. *et al.* Health and Population Effects of Rare Gene Knockouts in Adult Humans with Related Parents. *Science*, doi:10.1126/science.aac8624 (2016).

19 Lek, M. *et al.* Analysis of Protein-Coding Genetic Variation in 60,706 Humans. *Nature* **536**, 285-291, doi:10.1038/nature19057 (2016).

20 Kelly, M. P. *et al.* Phosphodiesterase 11a in Brain Is Enriched in Ventral Hippocampus and Deletion Causes Psychiatric Disease-Related Phenotypes. *Proc Natl Acad Sci U S A* **107**, 8457-8462, doi:10.1073/pnas.1000730107 (2010).

21 Di Angelantonio, E. *et al.* Lipid-Related Markers and Cardiovascular Disease Prediction. *Jama* **307**, 2499-2506, doi:10.1001/jama.2012.6571 (2012).

22 White, H. D. *et al.* Darapladib for Preventing Ischemic Events in Stable Coronary Heart Disease. *N Engl J Med* **370**, 1702-1711, doi:10.1056/NEJMoa1315878 (2014).

23 O'Donoghue, M. L. *et al.* Effect of Darapladib on Major Coronary Events after an Acute Coronary Syndrome: The Solid-Timi 52 Randomized Clinical Trial. *Jama* **312**, 1006-1015, doi:10.1001/jama.2014.11061 (2014).

24 Polfus, L. M., Gibbs, R. A. & Boerwinkle, E. Coronary Heart Disease and Genetic Variants with Low Phospholipase A2 Activity. *N Engl J Med* **372**, 295-296, doi:10.1056/NEJMc1409673 (2015).

25 Carr, B. A., Wan, J., Hines, R. N. & Yost, G. S. Characterization of the Human Lung Cyp2f1 Gene and Identification of a Novel Lung-Specific Binding Motif. *J Biol Chem* **278**, 15473-15483, doi:10.1074/jbc.M300319200 (2003).

26 Standiford, T. J. *et al.* Interleukin-8 Gene Expression by a Pulmonary Epithelial Cell Line. A Model for Cytokine Networks in the Lung. *J Clin Invest* **86**, 1945-1953, doi:10.1172/JCI114928 (1990).

27 Barnes, P. J. Chronic Obstructive Pulmonary Disease. *N Engl J Med* **343**, 269-280, doi:10.1056/NEJM200007273430407 (2000).

28 Murray, I. A., Coupland, K., Smith, J. A., Ansell, I. D. & Long, R. G. Intestinal Trehalase Activity in a Uk Population: Establishing a Normal Range and the Effect of Disease. *Br J Nutr* **83**, 241-245 (2000).

29 Christiansen, D. *et al.* Humans Lack Igb3 Due to the Absence of Functional Igb3-Synthase: Implications for Nkt Cell Development and Transplantation. *PLoS Biol* **6**, e172, doi:10.1371/journal.pbio.0060172 (2008).

30 Dahl, K., Buschard, K., Gram, D. X., d'Apice, A. J. & Hansen, A. K. Glucose Intolerance in a Xenotransplantation Model: Studies in Alpha-Gal Knockout Mice. *APMIS* **114**, 805-811, doi:10.1111/j.1600-0463.2006.apm_393.x (2006).

31 Casu, A. *et al.* Insulin Secretion and Glucose Metabolism in Alpha 1,3-Galactosyltransferase Knock-out Pigs Compared to Wild-Type Pigs. *Xenotransplantation* **17**, 131-139, doi:10.1111/j.1399-3089.2010.00572.x (2010).

32 Schneider, M. R. & Wolf, E. The Epidermal Growth Factor Receptor Ligands at a Glance. *J Cell Physiol* **218**, 460-466, doi:10.1002/jcp.21635 (2009).

33 Wang, G. X. *et al.* The Brown Fat-Enriched Secreted Factor Nrg4 Preserves Metabolic Homeostasis through Attenuation of Hepatic Lipogenesis. *Nat Med* **20**, 1436-1443, doi:10.1038/nm.3713 (2014).

34 Murtazina, R. *et al.* Tissue-Specific Regulation of Sodium/Proton Exchanger Isoform 3 Activity in Na(+)/H(+) Exchanger Regulatory Factor 1 (Nherf1) Null Mice. Camp Inhibition Is Differentially Dependent on Nherf1 and Exchange Protein Directly Activated by Camp in Ileum Versus Proximal Tubule. *J Biol Chem* **282**, 25141-25151, doi:10.1074/jbc.M701910200 (2007).

35 Wang, B., Yang, Y. & Friedman, P. A. Na/H Exchange Regulatory Factor 1, a Novel Akt-Associating Protein, Regulates Extracellular Signal-Regulated Kinase Signaling through a B-Raf-Mediated Pathway. *Mol Biol Cell* **19**, 1637-1645, doi:10.1091/mbc.E07-11-1114 (2008).

36 Karim, Z. *et al.* Nherf1 Mutations and Responsiveness of Renal Parathyroid Hormone. *N Engl J Med* **359**, 1128-1135, doi:10.1056/NEJMoa0802836 (2008).

37 Huff, M. W. & Hegele, R. A. Apolipoprotein C-Iii: Going Back to the Future for a Lipid Drug Target. *Circ Res* **112**, 1405-1408, doi:10.1161/CIRCRESAHA.113.301464 (2013).

38 Pollin, T. I. *et al.* A Null Mutation in Human Apoc3 Confers a Favorable Plasma Lipid Profile and Apparent Cardioprotection. *Science* **322**, 1702-1705, doi:10.1126/science.1161524 (2008).

39 Gaudet, D. *et al.* Antisense Inhibition of Apolipoprotein C-Iii in Patients with Hypertriglyceridemia. *N Engl J Med* **373**, 438-447, doi:10.1056/NEJMoa1400283 (2015).

40 Gaudet, D. *et al.* Targeting Apoc3 in the Familial Chylomicronemia Syndrome. *N Engl J Med* **371**, 2200-2206, doi:10.1056/NEJMoa1400284 (2014).

41 Graham, M. J. *et al.* Antisense Oligonucleotide Inhibition of Apolipoprotein C-Iii Reduces Plasma Triglycerides in Rodents, Nonhuman Primates, and Humans. *Circ Res* **112**, 1479-1490, doi:10.1161/circresaha.111.300367 (2013).

42 Brown, S. D. & Moore, M. W. Towards an Encyclopaedia of Mammalian Gene Function: The International Mouse Phenotyping Consortium. *Dis Model Mech* **5**, 289-292, doi:10.1242/dmm.009878 (2012).

43 Liu, Y. *et al.* Quantitative Variability of 342 Plasma Proteins in a Human Twin Population. *Mol Syst Biol* **11**, 786, doi:10.15252/msb.20145728 (2015).

44 Scott, E. M. *et al.* Characterization of Greater Middle Eastern Genetic Variation for Enhanced Disease Gene Discovery. *Nat Genet*, doi:10.1038/ng.3592 (2016).

45 Kooner, J. S. *et al.* Genome-Wide Association Study in Individuals of South Asian Ancestry Identifies Six New Type 2 Diabetes Susceptibility Loci. *Nat Genet* **43**, 984-989, doi:10.1038/ng.921 (2011).

46 Purcell, S. *et al.* Plink: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet* **81**, 559-575, doi:10.1086/519795 (2007).

47 Tennessen, J. A. *et al.* Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. *Science* **337**, 64-69, doi:10.1126/science.1219240 (2012).

48 Do, R. *et al.* Exome Sequencing Identifies Rare Ldlr and Apoa5 Alleles Conferring Risk for Myocardial Infarction. *Nature* **518**, 102-106, doi:10.1038/nature13917 (2015).

49 Fisher, S. *et al.* A Scalable, Fully Automated Process for Construction of Sequence-Ready Human Exome Targeted Capture Libraries. *Genome Biol* **12**, R1, doi:10.1186/gb-2011-12-1-r1 (2011).

50 Li, H. & Durbin, R. Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

51 McKenna, A. *et al.* The Genome Analysis Toolkit: A Mapreduce Framework for Analyzing Next-Generation DNA Sequencing Data. *Genome Res* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

52 DePristo, M. A. *et al.* A Framework for Variation Discovery and Genotyping Using Next-Generation DNA Sequencing Data. *Nat Genet* **43**, 491-498, doi:10.1038/ng.806 (2011).

53 Van der Auwera, G. A. *et al.* From Fastq Data to High Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Curr Protoc Bioinformatics* **11**, 11 10 11-11 10 33, doi:10.1002/0471250953.bi1110s43 (2013).

54 McLaren, W. *et al.* Deriving the Consequences of Genomic Variants with the Ensembl Api and Snp Effect Predictor. *Bioinformatics* **26**, 2069-2070, doi:10.1093/bioinformatics/btq330 (2010).

55 Jun, G. *et al.* Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data. *Am J Hum Genet* **91**, 839-848, doi:10.1016/j.ajhg.2012.09.004 (2012).

56 Manichaikul, A. *et al.* Robust Relationship Inference in Genome-Wide Association Studies. *Bioinformatics* **26**, 2867-2873, doi:10.1093/bioinformatics/btq559 (2010).

57 Gold, L. *et al.* Aptamer-Based Multiplexed Proteomic Technology for Biomarker Discovery. *PLoS One* **5**, e15004, doi:10.1371/journal.pone.0015004 (2010).

58 Hunter-Zinck, H. *et al.* Population Genetic Structure of the People of Qatar. *Am J Hum Genet* **87**, 17-25, doi:10.1016/j.ajhg.2010.05.018 (2010).

59 Purcell, S. M. *et al.* A Polygenic Burden of Rare Disruptive Mutations in Schizophrenia. *Nature* **506**, 185-190, doi:10.1038/nature12975 (2014).

60 Wright, S. Coefficients of Inbreeding and Relationship. *Am Nat* **56**, 330-338 (1922).

61 Price, A. L. *et al.* Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies. *Nat Genet* **38**, 904-909, doi:10.1038/ng1847 (2006).

62 Sambrook, J. & Russell, D. W. Purification of Nucleic Acids by Extraction with Phenol:Chloroform. *CSH Protoc* **2006**, doi:10.1101/pdb.prot4455 (2006).

63 Mosteller, R. D. Simplified Calculation of Body-Surface Area. *N Engl J Med* **317**, 1098, doi:10.1056/NEJM198710223171717 (1987).

64 Maraki, M. *et al.* Validity of Abbreviated Oral Fat Tolerance Tests for Assessing Postprandial Lipemia. *Clin Nutr* **30**, 852-857, doi:10.1016/j.clnu.2011.05.003 (2011).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements Dr. Saleheen is supported by grants from the National Institutes of Health, the Fogarty International, the Wellcome Trust, the British Heart Foundation, and Pfizer. Dr. Natarajan is supported by the John S. LaDue Memorial Fellowship in Cardiology from Harvard Medical School. Dr. Won is supported by a grant from the Samsung Medical Center, Korea (SMO116163). Dr. Kathiresan is supported by the Ofer and Shelly Nemirovsky MGH Research Scholar Award and by grants from the National Institutes of Health (R01HL107816), the Donovan Family Foundation, and Fondation Leducq. Exome sequencing was supported by a grant from the NHGRI (5U54HG003067-11) to Drs. Gabriel and Lander. Dr. MacArthur is supported by a grant from the National Institutes of Health (R01GM104371). In recognition for PROMIS fieldwork and support, we also acknowledge contributions made by the following: Mohammad Zeeshan Ozair, Usman Ahmed, Abdul Hakeem, Hamza Khalid, Kamran Shahid, Fahad Shuja, Ali Kazmi, Mustafa Qadir Hameed, Naeem Khan, Sadiq Khan, Ayaz Ali, Madad Ali, Saeed Ahmed, Muhammad Waqar Khan, Muhammad Razaq Khan, Abdul Ghafoor, Mir Alam, Riazuddin, Muhammad Irshad Javed, Abdul Ghaffar, Tanveer Baig Mirza, Muhammad Shahid, Jabir Furqan, Muhammad Iqbal Abbasi, Tanveer Abbas, Rana Zulfiqar, Muhammad Wajid, Irfan Ali, Muhammad Ikhlaq, Danish Sheikh, Muhammad Imran, Matthew Walker, Nadeem Sarwar, Sarah Venorman, Robin Young, Adam Butterworth, Hannah Lombardi, Binder Kaur and Nasir Sheikh. Fieldwork in the PROMIS study has been supported through funds available to investigators at the Center for Non-Communicable Diseases, Pakistan and the University of Cambridge, UK.

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

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