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

Danish Saleheen, Pradeep Natarajan, Irina M. Armean, Wei Zhao, Asif Rasheed, Sumeet A. Khetarpal, Hong-Hee Won, Konrad J. Karczewski, Anne H. O’Donnell-Luria, Kaitlin E. Samocha, Benjamin Weisburd, Namrata Gupta, Mozammal Zaidi, Maria Samuel, Atif Imam, Shahid Abbas, Faisal Majeed, Madiha Ishaq, Saba Akhtar, Kevin Trindade, Megan Mucksavage, Nadeem Qamar, Khan Shah Zaman, Zia Yaqoob, Tahir Saghri, Syed Nadeem Hasan Rizvi, Anis Memon, Nadeem Hayyat Mallick, Mohammad Ishaq, Syed Zahed Rasheed, Fazal-ur-Rehman Memon, Khalid Mahmood, Naveeduddin Ahmed, Ron Do, Ronald M. Krauss, Daniel G. MacArthur, Eric S. Lander, Mark J. Daly, Philippe Frossard, John Danesh, Daniel J. Rader, Namrata Gupta, Naveeduddin Ahmed, Daniel G. MacArthur, Eric S. Lander, Mark J. Daly, Philippe Frossard, John Danesh, Daniel J. Rader

A major goal of biomedicine is to understand the function of every gene in the human genome. Loss-of-function mutations can disrupt both copies of a given gene in humans and phenotypic analysis of such ‘human knockouts’ can provide insight into gene function. Consanguineous unions are more likely to result in offspring carrying homozygous loss-of-function mutations. In Pakistan, consanguinity rates are notably high. Here we sequence the protein-coding regions of 10,503 adult participants in the Pakistan Risk of Myocardial Infarction Study (PROMIS), designed to understand the determinants of cardiometabolic diseases in individuals from South Asia. We identified individuals carrying homozygous predicted loss-of-function (pLoF) mutations, and performed phenotypic analysis involving more than 200 biochemical and disease traits. We enumerated 49,138 rare (\(< 1\%\) minor allele frequency) pLoF mutations. These pLoF mutations are estimated to knock out 1,317 genes, each in at least one participant. Homozygosity for pLoF mutations at PLA2G7 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 NR4G, with markedly reduced plasma insulin C-peptide concentrations; and at SLC9A3R1, with mediators of calcium and phosphate signalling. Heterozygous deficiency of APOC3 has been shown to protect against coronary heart disease; we identified APOC3 homozygous pLoF carriers in our cohort. We recruited these human knockouts and challenged them with an oral fat load. Compared with family members lacking the mutation, individuals with APOC3 knocked out 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.

Across all participants (Table 1), exome sequencing yielded 1,639,223 exonic and splice-site sequence variants in 19,026 autosomal genes that passed initial quality control metrics. Of these, 57,137 mutations across 14,345 autosomal genes were annotated as pLoF mutations (that is, nonsense, frameshift, or canonical splice-site mutations predicted to inactivate a gene). To increase the probability that mutations are correctly annotated as pLoF by automated algorithms, 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. The site-frequency spectrum for these pLoF mutations revealed that the majority was seen only in one or a few individuals (Extended Data Fig. 1).

Across all 10,503 PROMIS participants, both copies of 1,317 distinct genes were predicted to be inactivated owing 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 of the participants that were sequenced (1,843 individuals, 17.5%) had at least one gene knocked out by a homozygous pLoF mutation. 1,504 of these 1,843 individuals (81.6%) were homozygous pLoF carriers for just one gene, but the minority of participants had more than one gene knocked out 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 outbred ancestor. PROMIS participants had a fourfold higher median inbreeding coefficient compared to outbred populations (0.016 versus 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...
Table 1 | Baseline characteristics of exome sequenced study participants

| Characteristic                  | Value (n = 10,503) |
|---------------------------------|--------------------|
| Age (years), mean (s.d.)        | 52.0 (9.0)         |
| Women, number (%)               | 1,802 (17.2%)      |
| Parents closely related, number (%) | 4,101 (39.0%)    |
| Spouse closely related, number (%) | 4,182 (39.8%)    |
| Ethnicity, number (%)           |                    |
| 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, number (%)        | 4,744 (45.2%)      |
| Hypercholesterolemia, number (%)| 2,924 (27.8%)      |
| Diabetes mellitus, number (%)   | 4,264 (40.6%)      |
| Coronary heart disease, number (%)| 4,793 (45.6%)    |
| Smoking, number (%)             | 4,201 (40.0%)      |
| BMI (kg·m$^{-2}$), mean (s.d.)  | 25.9 (4.2)         |

1Hypertension defined as systolic blood pressure ≥ 140 mm Hg, diastolic blood pressure ≥ 90 mm Hg, or antihypertensive treatment.
2Hypercholesterolemia defined as serum total cholesterol >240 mg d$^{-1}$, or lipid lowering therapy or self-report.
3Diabetes defined as fasting glucose ≥ 126 mg d$^{-1}$, or HbA1c > 6.5%, oral hypoglycemics, insulin treatment, or self-report.
4Coronary heart disease defined as acute myocardial infarction as determined by clinical symptoms with typical EKG findings or elevated serum troponin I.
5Smoking defined as active current or prior tobacco smoking.

Lipoprotein-associated phospholipase A2 (Lp-PLA2, encoded by PLA2G7) hydrolyses phospholipids to generate lysophosphatidylcholine and oxidized non-esterified 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\(^6\). 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 + 1 G>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 + 1 G>A homozygotes have markedly lower Lp-PLA2 enzymatic activity (−245 nmol ml$^{-1}$ min$^{-1}$, P = 2 × 10$^{-7}$), whereas the 106 heterozygotes had an intermediate effect (−120 nmol ml$^{-1}$ min$^{-1}$, P = 2 × 10$^{-7}$) (Fig. 2a, b). If Lp-PLA2 has a causal role for coronary heart disease, one might expect those naturally that are deficient for this enzyme to have reduced risk for coronary heart disease. We tested the association of PLA2G7 c.663 + 1 G>A with myocardial infarction across all participants and found that carriers of the pLoF allele did not have reduced risk\(^{10,11}\) (odds ratio 0.97; 95% confidence interval, 0.70–1.34; P = 0.87) (Fig. 2c). In contrast, at two positive control genes, we replicated previous observations (Supplementary Table 4); at LDLR, homozygous pLoF mutations increased myocardial infarction risk by 20-fold and, at PCSK9, heterozygous pLoF mutations reduced risk by 78%. Note, in two recent randomized controlled trials, pharmacologic Lp-PLA2 inhibition failed to reduce risk for coronary heart disease\(^{12,13}\), a result that might have been anticipated by this genetic analysis.

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\(^{14}\). At CYP2F1, we identified two participants homozygous for a splice-site mutation, c.1295-2(A>G). When compared with non-carriers, c.1295-2(A>G) homozygotes displayed higher soluble interleukin-8 (IL-8) concentrations (3.7-fold increase, P = 2 × 10$^{-6}$) (Extended Data Fig. 4). CYP2F1 c.1295-2(A>G) heterozygosity had a more modest effect (2.4-fold increase, P = 2 × 10$^{-4}$). IL-8 induces migration of neutrophils in airways and is a mediator of acute pulmonary inflammation and chronic obstructive pulmonary disease (COPD)\(^{15}\). However, neither of the carriers reported a personal or family history of obstructive pulmonary disease; further studies of these participants are required to assess the roles of CYP2F1 and IL-8 on pulmonary physiology.

Trehalase (encoded by TREH) is an intestinal enzyme that splits the naturally occurring unabsorbed disaccharide trehalose into two glucose molecules\(^{16}\). 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.90-9106 deletion 5’-TCTCTGCAG TTAGATTACTGCCCAGC-3’) in exon 2. Homozygotes, unlike heterozygotes or non-carriers, had lower concentrations of several apolipoprotein B-containing lipoprotein subfractions (Supplementary Table 3, Extended Data Fig. 5).

α1,3-galactosyltransferase 2 (encoded by A3GALT2) catalyses the formation of the Gal-α1,3Galβ1,4GlcNAc-R (α-gal) epitope; the biological role of this enzyme in humans is uncertain\(^{17}\). At A3GALT2, we identified two participants homozygous for a frameshift mutation, p.Thr106SerfsTer4. Compared with non-carriers, p.Thr106SerfsTer4 homozygotes both had markedly reduced concentrations of fasting C-peptide (−97.4%; P = 6 × 10$^{-12}$) and insulin (−92.3%; P = 1 × 10$^{-8}$). Such an association was only observed in the homozygous state (Extended Data Fig. 6). Interestingly, A3galt2−/− mice and pigs have recently been shown to have glucose intolerance\(^{8,19}\).

To understand whether the identification of only a single homozygote may still be informative, we performed a complementary analysis, focused on those with the most extreme standard Z scores (|Z score| > 5)}
and with the requirement that there was also evidence for association in heterozygotes (see Methods). This procedure highlighted neuregulin 4 (NRG4), a member of the epidermal growth factor family extracellular ligands that is highly expressed in brown fat, particularly during adipocyte differentiation\textsuperscript{20,21}. 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 a 48.3% reduction in insulin C-peptide (\( P = 1 \times 10^{-2} \)). Mice in which Nrg4 is deleted have recently been shown to have glucose intolerance\textsuperscript{22}. 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 pheno-
typing will be required to definitively assess any relationship of NRG4 deficiency in humans with glucose intolerance.

To further dissect the effects 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; we associated these genotypes across 1,310 protein biomarkers and observed a number of associations (Supplementary Table 5). We highlight two PROMIS participants with 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 5). SLC9A3R1 encodes a Na\(^+\)/H\(^+\) exchanger regulatory cofactor that interacts with and regulates the parathyroid hormone receptor; Slec9a3r1\(^{−/−}\) mice display hyperphosphaturia and disrupted protein-kinase-A-dependent cAMP-mediated phosphorylation\textsuperscript{22}. Humans carrying rare missense mutations in SLC9A3R1 have nephro lithiasis, osteoporosis, and hypophosphatemia\textsuperscript{23}.

Apolipoprotein C3 (apoC3, encoded by APOC3) is a major protein component of chylomicrons, very low-density lipoprotein cholesterol, and high-density lipoprotein cholesterol\textsuperscript{24}. We and others recently reported that heterozygous APOC3 pLoF mutations lower plasma triglycerides and reduce risk for coronary heart disease\textsuperscript{26,27}; there is now substantial interest in APOC3 as a therapeutic target\textsuperscript{26–28}. In published studies, no APOC3 pLoF homozygotes have been identified despite study of nearly 200,000 participants from the US and Europe, raising concerns that complete APOC3 deficiency may be harmful. However, in this study of around 10,000 Pakistanis, we identified four participants homozygous for APOC3 p.Arg19Ter. When compared with non-carriers, p.Arg19Ter homozygotes displayed near-absent plasma apoC3 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 \( \text{dl}^{-1} \), \( P = 3 \times 10^{-5} \)); and similar levels of low-density lipoprotein (LDL) cholesterol (\( P = 0.14 \)) (Fig. 3a–d).

ApoC3 functions as a brake on the clearance of dietary fat from the circulation 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 were surprised to find that the wife of the proband, a first cousin, was also a pLoF homozygote, leading to all nine children being obligate homozygotes (Fig. 3e). In this family, we challenged pLoF homozygous carriers (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 \( \text{dl}^{-1}\) over 6 h (AFR)) and European (EUR) ancestry. c. The burden of homozygous pLoF genes per individual is correlated with coefficient of inbreeding. Bars represent 1.5× interquartile range beyond the 25th and 75th percentiles (b, c).

**Figure 1** | Homozygous pLoF burden in PROMIS is driven by excess autozygosity. 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 and European (EUR) ancestry.

**Figure 2** | Carriers of PLA2G7 splice mutation have diminished Lp-PLA2 mass and activity but similar risk for coronary heart disease when compared to non-carriers. 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 (\( P = 6 \times 10^{-4} \)) and activity (\( P = 2 \times 10^{-7} \)), 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 (\( P = 0.87 \)). Bars represent 1.5× interquartile range beyond the 25th and 75th percentiles (a, b).
versus 1,267.7 mg dl⁻¹ over 6 h; P = 1 × 10⁻⁴) (Fig. 3f). These data show that complete lack of apoC3 markedly improves clearance of plasma triglycerides after a fatty meal and are consistent with and extend an earlier report of diminished post-prandial lipaemia in APOC3 pLoF heterozygotes.

Targeted gene disruption in model organisms followed by phenotypic analysis has been a fruitful approach in understanding gene function; here, we extend this concept to humans, leveraging naturally occurring pLoF mutations, consanguinity, and biochemical phenotyping. Our 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% confidence interval, 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, re-contact by genotype followed by provocative testing may provide physiologic insights. We used this approach to demonstrate that complete lack of apoC3 is tolerated and results in both lowered fasting triglyceride concentrations as well as substantially blunted post-prandial lipaemia.

Several limitations deserve mention. First and most importantly, any given mutation annotated as pLoF may not truly lead to loss of protein function. To address this issue, in addition to bioinformatics filtration, we performed manual curation on all homozygous pLoF variants (n = 1,580) (Supplementary Tables 1, 6). Of note, such manual curation was not described in earlier reports. We found 56 variants leading to nine obligate homozygote children. Given the extensive number of first-degree unions, the pedigree is simplified for clarity. 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. Bars represent 1.5 × interquartile range beyond the 25th and 75th percentiles (a–d, f).

Figure 3 | APOC3 pLoF homozygotes have diminished fasting triglycerides and blunted post-prandial lipaemia. a–d. APOC3 pLoF genotype status, apolipoprotein C3, triglycerides, HDL cholesterol and LDL cholesterol distributions among all sequenced participants. Apolipoprotein C3 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 number of 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. Bars represent 1.5 × interquartile range beyond the 25th and 75th percentiles (a–d, f).

Figure 4 | Simulations anticipate many more homozygous pLoF genes in the PROMIS cohort. 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.
with genotypes with a low number of supportive reads, 55 with poorly mapped reads (Supplementary Table 7), and an additional 66 in which there were potential mechanisms of protein-truncation rescue (Extended Data Fig. 7) or occurred within exons or splice sites at which conservation was low. Thus, we found that the majority of pLoF calls (1,403 out of 1,580; 89%) were free of mapping or annotation error. However, for any given pLoF, experimental validation will be required to prove loss of gene function (for example, targeted assays such as reverse-transcription PCR of transcript and/or western blot of protein to confirm its absence in the relevant tissue). A second limitation is reduced statistical power for genotype–phenotype correlation if a gene is knocked out in only one or two participants. However, this could be improved with larger sample sizes (Extended Data Fig. 8). Finally, our analysis was limited to available genotypes and in only one instance did we recall participants for deeper phenotyping; rather, a standardized clinical phenotyping protocol is desirable for each participant in which a gene is observed to be knocked out.

These observations pave the way 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 in which homozygous genotypes may be enriched; (2) deep-coverage sequencing of the protein-coding regions of the genome; (3) availability of a broad array of biochemical as well as clinical phenotypes across the population; (4) ability to re-contact knockout families and their follow-up patients; (5) a thorough clinical evaluation in each participant in which a gene is observed to be knocked out; and (6) hypothesis-driven provocative phenotyping in selected participants.

Online Content Methods, along with any additional Extended Data items and Source Data, are available in the online version of this journal. References unique to these sections appear only in the online version of the paper.

Received 22 October 2015; accepted 5 March 2017.

1. Eisenberg, D., Marcotte, E. M., Xenarios, I. & Yeates, T. O. Protein function in the post-genomic era. Nature 405, 823–826 (2000).
2. Bittles, A. H., Mason, W. M., Greene, J. & Rao, N. A. Reproductive selected participants.
3. Narasimhan, V. M.
4. Gregson, J. M.
5. Metzger, B. J., Selvin, S. R., Carson, A. L. & Haffner, S. M. Glucose intolerance in a younger and more diverse cohort: results from the alpha-gal knockout mice. PLoS Biol. 6, e172 (2008).
6. Dahl, K., Buschard, K., Gram, D. X., d’Apice, A. J. & Hansen, A. K. Glucose intolerance in a younger and more diverse cohort: results from the alpha-gal knockout mice. PLoS Biol. 6, e172 (2008).
7. Sulem, P.
8. Casu, A. et al. Insulin secretion and glucose metabolism in alpha 1,3-galactosyltransferase knockout-pigs compared to wild-type pigs. Xenotransplantation 17, 1–139 (2010).
9. Schneider, M. R. & Wolf, E. The epidermal growth factor receptor ligands at a glance. J. Cell. Physiol. 218, 460–466 (2008).
10. Wang, G. X. et al. The brown fat-enriched secreted factor Fgf4 preserves metabolic homeostasis through attenuation of hepatic lipogenesis. Nat. Med. 20, 1436–1443 (2014).
11. Tung, T. M. et al. A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent cardiovascular protection. Science 322, 1702–1705 (2008).
12. Gaudet, D. et al. Antisense inhibition of apolipoprotein C-III in patients with hypertriglyceridermia. N. Engl. J. Med. 373, 438–447 (2015).
13. Gaudet, D. et al. Targeting APOC3 in the familial chylomicronemia syndrome. N. Engl. J. Med. 371, 2200–2206 (2014).
14. Graham, M. J. et al. Antisense oligonucleotide inhibition of apolipoprotein C-III reduces plasma triglycerides in rodent, nonhuman primates, and humans. Circ. Res. 112, 1479–1490 (2013).
15. Brown, S. D. & Moore, M. W. Towards an encyclopedia of mammalian gene function: the International Mouse Phenotyping Consortium. Dis. Model. Mech. 5, 289–292 (2012).
16. Scott, E. M. et al. Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. Nat. Genet. 48, 1071–1076 (2016).

Supplementary Information is available in the online version of the paper.

Acknowledgements D.S. is supported by grants from the National Institutes of Health, the Fogarty International, the Wellcome Trust, the British Heart Foundation, and Pfizer. P.N. is supported by the John S. LaDue Memorial Fellowship in Cardiology from Harvard Medical School. H.-W.H. is supported by a grant from the Samsung Medical Center, Korea (SMO116163). S.K. is supported by the Ofer and Shelly Nemirovsky MGH Research Scholar Award and by grants from the National Institutes of Health (HL107816), the Donor Family Foundation, and Foundation Leducq. Exome sequencing was supported by a grant from the NHGRI (5U54HG003067-1) and to S.G. and E.S.L. D.G.M. is supported by a grant from the National Institutes of Health (R01GM104371). J.D. holds a British Heart Foundation Chair, European Research Council Senior Investigator Award, and NHR Senior Investigator Award. The Cardiovascular Epidemiology Unit at the University of Cambridge, which supported the field work and genotyping of PROMIS, is funded by the UK Medical Research Council, British Heart Foundation, and NHRC Cambridge Epidemiological Research Centre. In addition to their direct support, we also acknowledge contributions made by the following: M. Z. Ozair, U. Ahmed, A. Hakeem, H. Khalid, K. Shahid, F. Shuja, A. Kazmi, M. Qadir Hameed, N. Khan, S. Khan, A. Ali, M. Ali, S. Ahmad, M. W. Khan, M. R. Khan, A. Ghaffoor, M. Alam, R. Ahmed, M. J. Javed, A. Ghaffir, T. B. Mirza, M. Shahid, J. Furqan, M. I. Abbasi, T. Abbass, R. Zulfiquar, M. Wajid, I. Ali, M. Iklas, D. Sheikh, M. Imran, M. Walker, S. Narwar, S.Venorman, R. Young, A. Butterworth, H. Lombardi, B. Kaur and N. 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.D., J.F., D.R., M.Z., A.M., S.A., J.A., T.H., K.M., K.Z., R.M.K., D.S., P.F., J.D., and W.Z. performed array-based genotyping and runs-of-homoygosity analyses. Exome sequencing was coordinated by D.S., N.G., S.G., E.L., D.J.R., and S.K.P., N.W., H.H.W., and R.D. performed exome-sequencing quality control and association analyses. P.N., I.M.A., K.K.J., A.H.O., B.W., 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., K.M., K.Z., and M.I. performed constraint score analyses. D.S., P.N., and S.K. designed the study and wrote the paper. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to S.K. (sekar@broadinstitute.org) or D.S. (saleheen@mail.med.upenn.edu).

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.
Methods
Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

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 centres 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−1 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 min of venipuncture. Serum, plasma and whole blood samples were stored at −70°C within 45 min 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 analysers. 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 turbidimetric assay in whole-blood samples (Roche Diagnostics, USA). For further measurements, aliquots of serum and plasma samples were transported on dry ice to the Smillow Research Center, University of Pennsylvania, USA, where following biochemical analyses were conducted: apolipoproteins (apoA1, apoA2, apoB, apoC3, apoE) and non-esterified fatty acids were measured through immunoturbidimetric assays using kits by Roche Diagnostics or Kamiya; lipoprotein (a) levels were determined through a turbidimetric assay using reagents and calibrators from Denka Seiken; LpPLA2 mass and activity levels were determined using immunoassays manufactured by diadexus; measurements for insulin, leptin and adiponectin were made using radio-immunoassays by LINCO; 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, chloride, phosphate, sex-hormone binding globulin were measured using enzymatic assays manufactured by Abbott Diagnostics. Glomerular filtration rate (eGFR) was estimated from serum creatinine levels using the MDRD equation. apoC3 levels were determined in an autoanalyser using a commercially available ELISA by Sekisui Diagnostics. 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): fatty acid binding protein, granulocyte-macrophage colony stimulating factor, granulocyte colony stimulating factor, IL-1α, 1β, 1β, 2, 3, 4, 5, 6, 7, 10, 11, 12, 13, 18, 19, 20, 23, 24, 28, 29, 31, 33, α1acid glycoprotein, matrix metalloproteinase 2, matrix metalloproteinase 3, matrix metalloproteinase 9, nerve growth factor (β), tumour necrosis factor α, tumour necrosis factor β, 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, strong cell surface receptor 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. Initial quality control 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 < 10−5 for the Hardy–Weinberg equilibrium test were also excluded from the analyses. In PROMIS, further quality control included removal of participants with discrepancy between their reported sex and genetic sex determined from the X chromosome. To identify sample duplicates, 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. Laboratory methods for exome sequencing. Exome sequencing was performed at the Broad Institute. Sequencing and exome capture methods have been previously described. 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, Massachusetts, USA). DNA concentration was determined by PicoGreen (Invitrogen) before storage in 2-D barcoded 0.75 ml Matrix tubes at −20°C in the SmaRTstore (RTS,b Chester, UK) automated sample handling system. Initial quality control on all samples involving sample quantification (PicoGreen), confirmation of high-molecular weight DNA and fingerprint genotyping and gender determination (Illumina Iselect; Illumina). 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, with the following modifications: initial genomic DNA input into shearing was reduced from 3 μg to 10–100 ng in 50 μl 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-addition, adapter ligation, and library enrichment PCR were purchased from KAPA Biosciences (Wilmington, Massachusetts, 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 were used for in-solution hybrid selection as previously described, with the following exception: before 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 before hybridization, and second, the Midl 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 2 nM and pooled by equal volume using the Hamilton Starlet. Pools were then denatured using 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 analysed using RTA v1.12.4.2. Each pool of whole-exome libraries was run on paired 76-bp read 1 and 2. Libraries for end repair, KAPA library enrichment PCR 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 raw-time basecalls (RTA v1.12.4.2 software (Bustard), converted to .seq.txt files, and aligned to a human reference (hg19) using Burrows–Wheelier Aligner (BWA)®). 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®. 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 mpileupCaller tool from the GATK to generate a vCF (variant call format). Joint genotyping was subsequently performed and raw variant data for each sample was formatted (variant call format). Single-nucleotide polymorphism (SNPs) and indel sites were initially filtered after variant calibration marked sites of low quality that were likely false positives.

Data analysis quality control. 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 from historical values was noted.

Data processing and quality control of exome sequencing. Variant annotation. Variants were annotated using Variant Effect Predictor® and the LOFTEE® 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 SNPs 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 estimates the proportion of non-reference bases at reference sites, and excluded samples with high estimated contamination (FREEMIX scores >0.2)41. After removing monomorphic twins or duplicate samples using the KING software42, we removed outlier samples with too many or too few SNPs (>17,000 or <12,000 total variants; >400 singlonts; 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 >0.6 or <2). Finally, we removed samples with high missingness (>0.05).

Variant level quality control. Variant score quality recalibration was performed separately for SNPs and indels using the GATK VariantRecalibrator and ApplyRecalibration to filter out variants with lower accuracy scores. Additionally, we removed sites with an excess of heterozygosity calls (inbreeding coefficient < −0.3). To further reduce the rate of inaccurate variant calls, we further filtered out SNPs with low average quality (quality per depth of coverage <2) and a high degree of missing data (>20%), and indels also with low average quality (quality per depth of coverage <3) and a high degree of missing data (>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). Protein-capture was performed using modified aptamer technology as previously described43. In brief, 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 affiliation of isolation, or >5% of traits in the top or bottom 1st percentile of the analytic distribution.

Methods for manual curation 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 10.

Methods for inbreeding analyses. Array-derived runs of homozygosity. Analyses were conducted in PLINK44 using genome-wide association (GWAS) data in PROMIS and HapMap3 populations. Segments of the genome that were at least 1.5 Mb in length, had a SNP density of 1 SNP per 20 kb and had 25 consecutive homozygous SNPs (1 homozygous 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 previously45. 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 HapMap3 populations, respectively. To investigate variability in homozygosity explained by parental consanguinity, the difference in R2, or proportion of homozygosity variation explained by the model, 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.

In PROMIS, 39.0% of participants reported that their parents were cousins and 39.8% reported that they themselves were married to a cousin. An expectation from consanguinity is long regions of autozygosity, defined as homozygous loci identical by descent44. 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, Utah residents with Northern and Western European ancestry; TSI, Toscani in Italy) (P = 3.6 × 10−32), East Asian (CHB, Han Chinese in Beijing, China; JPT, Japanese in Tokyo, Japan; CHD= Chinese in Denver, Colorado) (P = 5.4 × 10−48) and African ancestries (YRI, Yoruban in Ibadan, Nigeria; MKK, Maasai in Kinyawa, Kenya) (P = 1.3 × 10−48), respectively (Extended Data Fig. 9).

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 had exome sequenced at the Broad Institute (Cambridge, Massachusetts) by the Myocardial Infarction Genes consortium46. We extracted approximately 5,000 high-quality polymorphic SNPs in linkage equilibrium present on both target intervals that passed variant quality control criteria based on HapMap3 data47. Using PLINK, we estimated the coefficient of inbreeding separately within each ethnicity group48. The coefficient of inbreeding was estimated as the observed degree of homozygosity compared with the anticipated homozygosity derived from an estimated common ancestor49. 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 + F × CMAF, which accounts for allozygous and autozygous, respectively, mechanisms for complete gene 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 (Extended Data Fig. 10). For a range of sample sizes (0–200,000), each gene was randomly sampled under a binomial distribution (X ~ (n, CMAF)), where X is the carrier probability distribution and n is the number of individuals sequenced, 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 Consortium50,51. In brief, we used the number of observed and rare loss-of-function (pLoF) (MAF <0.1%) high-confidence loss-of-function mutations was calculated. 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% 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 χ2 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 appreciable pLoF allele frequencies yet relative depletion of homozygous pLoF genotypes. We computed estimated genotype frequencies on the basis of Hardy–Weinberg equilibrium and the F inbreeding coefficient and compared the frequencies to the observed genotype counts with the χ2 goodness-of-fit test. A nominal P <0.05 is used to demonstrate at least nominal association.

The observed 1,317 homozygous pLoF genes were less likely to be classified as highly constrained (odds ratio 0.14; 95% confidence interval, 0.12, 0.16; P < 1 × 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%)52.
A number of genes previously predicted to be required for viability in humans were observed in the homozygous pLoF state in humans (Supplementary Table 8). For example, 40 of the 1,317 genes were found to be associated with embryonic or perinatal lethality as homozygous pLoF in mice21. Furthermore, 56 genes predicted to be essential using mouse/human conservation data25 are tolerated as homozygous pLoF in Pakistani adults. In fact, nine genes are in both datasets and are also modelled as loss-of-function intolerant21. One such gene, EP400 (also known as p400), influences cell cycle regulation via chromatin remodelling26 and is critical for maintaining the identity of murine embryonic stem cells25 but we observe an adult human homozygous for disruption of a canonical splice site (intron 3 of 52; c.1435 + 1 G>A) in EP400. Conversely, we observed 90 genes where the heterozygous pLoF genotype is of appreciable frequency but the homozygous pLoF genotype is depleted (P-value threshold <0.05) (Supplementary Table 9).

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 (MAF < 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 pairs. To reduce the likelihood of false positives, we only considered gene–trait pairs in which 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 analyses26. For lipoprotein-related traits, the use of lipid-lowering therapy was used as a covariate. For glycomic biomarkers, only non-diabetics were used in the analysis. The P threshold for statistical significance was 0.05/18,959 = 3 × 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 gene homozygotes. We performed an exploratory analysis of gene–trait pairs where there was only one 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 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 plates, age, sex, myocardial infarction status, and principal components. Gene–analyte associations were considered significant if P values were less than 0.05/(1,310 × 9) = 4.3 × 10−5. 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 members of his family and community members for DNA extraction and separated into plasma for lipid and apolipoprotein measurements. All subjects provided consent before initiation of the studies. IRB: 00007048 at the Center for Non-Communicable Diseases, Pakistan). DNA was isolated from whole blood using a reference phenol–chloroform protocol27. 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 5′-CTCGCTTTCGCGACGACCGTACGAAAG-3′, reverse primer 5′-CTTGAACGCTGCCGGGCGGAAAG-3′, PCR products were purified with Exo-SAP-IT (Affymetrix) and sequenced using Sanger sequencing using the same primers.

Oral fat tolerance test. Six non-carriers and seven homozygous carriers also participated in an oral fat tolerance test. Participants fasted overnight and then blood was drawn for measurement of baseline fasted fats. 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 in ref. 58). Participants consumed this oral load within a time span of 20 min and afterwards consumed 200 ml of water. Blood was drawn at 2, 4, and 6 h after oral fat consumption as done previously59. All lipid and apolipoprotein measurements from these plasma samples were determined by immunoturbidimetric assays on an ACE Axcel Chemistry analyser (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.

Data availability. Summaries of all pLoF variants observed in a homozygous are in the Supplementary Information. They are additionally all observed protein-coding variation, publicly available in the Exome Aggregation Consortium browser (http://exac.broadinstitute.org). DNA sequences have been deposited with the NIH dbGAP repository under accession numbers phs000917.
Extended Data Figure 1 | pLoF mutations are typically seen in very few individuals. The site-frequency spectrum of synonymous, missense, and high-confidence pLoF mutations is represented. Points represent the proportion of variants within a $1 \times 10^{-4}$ minor allele frequency bin for each variant category. Lines represent the cumulative proportions of variants categories. The bottom inset highlights that most pLoF variants are often seen in no more than one or two individuals. The top inset highlights that virtually all pLoF mutations are very rare.
Extended Data Figure 2 | Intersection of homozygous pLoF genes between PROMIS and other cohorts. We compared the counts and overlap of unique homozygous pLoF genes in PROMIS with other exome sequenced cohorts.
Extended Data Figure 3 | QQ-plot of recessive model pLoF association analysis across phenotypes. Analyses to determine whether homozygous pLoF carrier status was associated with traits was performed where there were at least two homozygous pLoF carriers phenotyped per trait. The observed versus the expected results from 15,263 associations are displayed here demonstrating an excess of associations beyond a Bonferroni threshold.
Extended Data Figure 4 | Carriers of pLoF alleles in CYP2F1 have increased IL-8 concentrations. Participants who had pLoF mutations in the CYP2F1 gene had higher concentrations of IL-8, whereas heterozygotes had a more modest effect when compared to the rest of the cohort of non-carriers. IL-8 concentration is natural log transformed. Bars represent $1.5 \times$ interquartile range beyond the 25th and 75th percentiles.
Extended Data Figure 5 | Carriers of pLoF alleles in TREH have decreased concentrations of several lipoprotein subfractions. Participants who had pLoF mutations in the TREH gene had lower concentrations of several lipoprotein subfractions. Bars represent 1.5 × interquartile range beyond the 25th and 75th percentiles.
Extended Data Figure 6 | Nondiabetic homozygous pLoF carriers for A3GALT2 have diminished insulin C-peptide concentrations. Among nondiabetics, those who were homozygous pLoF for A3GALT2 had substantially lower fasting insulin C-peptide concentrations. This observation was not evident in nondiabetic heterozygous pLoF A3GALT2 participants. Insulin C-peptide is natural log transformed. Bars represent 1.5× interquartile range beyond the 25th and 75th percentiles.
Extended Data Figure 7  | Example of a second polymorphism in-phase which rescues a putative protein-truncating mutation. Short-reads that align to genomic positions 65,339,112 to 65,339,132 on chromosome 1 are displayed for one individual with a putative homozygous pLoF genotype in this region. The SNP at position 65,339,122 from G to T is annotated as a nonsense mutation in the JAK1 gene. However, all three homozygotes of this mutation carried a tandem SNP in the same codon (A to G at 65,339,124) thus resulting in a glutamine and effectively rescuing the protein-truncating mutation.
Extended Data Figure 8 | Anticipated number of genes knocked out with increasing sample sizes by minimum knockout count. We simulate the number of genes expected to be knocked out by minimum knockout count per gene at increasing sample sizes. We perform this simulation with and without the observed inbreeding.
Extended Data Figure 9 | PROMIS participants have an excess burden of runs of homozygosity compared with other populations. Consanguinity leads to regions of genomic segments that are identical by descent and can be observed as runs of homozygosity. Using genome-wide array data in 17,744 PROMIS participants and reference samples from the International HapMap3, the burden of runs of homozygosity (minimum 1.5 Mb) per individual was derived and population-specific distributions are displayed, with outliers removed. This highlights the higher median runs of homozygosity burden in PROMIS and the higher proportion of individuals with very high burdens.
Extended Data Figure 10 | Down-sampling of synonymous and high confidence pLoF variants to validate simulation. a, b. We ran simulations to estimate the number of unique, completely knocked out genes at increasing sample sizes. Before applying our model, we first applied this approach to a range of sample sizes below 7,078 for variants that were not under constraint, synonymous variants (a), and for high-confidence null variants (b). At the observed sample size, we did not observe significant selection. We expect that at increasing sample sizes, there may be a subset of genes that will not be tolerated in a homozygous pLoF state. In fact, our estimates are slightly more conservative when comparing outbred simulations with a recent description of >100,000 Icelanders using a more liberal definition for pLoF mutations.