LETTER

doi:10.1038/nature22034

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

Danish Saleheen[1,2]*, Pradeep Natarajan[3,4]*, Irina M. Armean[4,5], Wei Zhao[1], Asif Rasheed[2], Sumeet A. Khetarpal[6], Hong-Hee Won[7], Konrad J. Karczewski[4,5], Anne H. O'Donnell-Luria[4,5,8], Kaitlin E. Samocha[4,5], Benjamin Weisburd[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]§

A major goal of biomedicine is to understand the function of every gene in the human genome[1]. 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[2]. 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[3]. 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 *NRG4*, 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[4,5]; 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

[1]Department of Biostatistics and Epidemiology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pennsylvania, USA. [2]Center for Non-Communicable Diseases, Karachi, Pakistan. [3]Center for Genomic Medicine, Massachusetts General Hospital and Department of Medicine, Harvard Medical School, Boston, Massachusetts, USA. [4]Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. [5]Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA. [6]Institute for Translational Medicine and Therapeutics, Department of Genetics, and Department of Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pennsylvania, 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, Massachusetts, 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, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [17]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [18]Children's Hospital Oakland Research Institute, Oakland, California, USA. [19]MRC/BHF Cardiovascular Epidemiology Unit, 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.

*These authors contributed equally to this work.

§These authors jointly supervised this work.

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, nubmer (%)	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 (m kg^{-2}), mean (s.d.)	25.9 (4.2)

*Hypertension defined as systolic blood pressure ≥ 140 mm Hg, diastolic blood pressure ≥ 90 mm Hg, or antihypertensive treatment.
†Hypercholesterolemia defined as serum total cholesterol >240 mg dl^{-1}, lipid lowering therapy or self-report.
‡Diabetes defined as fasting blood glucose ≥ 126 mg dl^{-1}, 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.

those who did not (0.023 versus 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's $r = 0.31$; $P = 5 \times 10^{-231}$) (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 compared our results to three recent reports in which homozygous pLoF genes have been catalogued: in Pakistanis living in Britain[6], in Icelanders[7], and in the Exome Aggregation Consortium (ExAC)[8]. In the PROMIS study, we identify a total of 734 unique genes with homozygous pLoF mutations that were not observed in the other three studies (Extended Data Fig. 2). Intersection of the four sets of genes from these studies revealed only 25 common to all four studies.

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 at which two or more participants had homozygous pLoF mutations, we conducted an association screen against 201 distinct phenotypes (Supplementary Table 2). 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 C3 (encoded by *APOC3*), we recruited participants on the basis of genotype and performed provocative physiologic testing.

In an association screen of knockout genes with phenotypes, the quantile–quantile plot of expected versus observed association results shows an excess of highly significant results without systematic inflation (Extended Data Fig. 3). Association results surpassed the Bonferroni significance threshold ($P = 3 \times 10^{-6}$, see Methods) for 26 gene–trait pairs (Supplementary Table 3). Below, we highlight seven results: *PLA2G7*, *CYP2F1*, *TREH*, *A3GALT2*, *NRG4*, *SLC9A3R1*, and *APOC3*.

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[9]. 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}, $P = 2 \times 10^{-7}$), whereas the 106 heterozygotes had an intermediate effect (-120 nmol ml^{-1} min^{-1}, $P = 2 \times 10^{-77}$) (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*, heterozygous pLoF mutations increased myocardial infarction risk by 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[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, $P = 2 \times 10^{-6}$) (Extended Data Fig. 4). *CYP2F1* c.1295-2(A > G) heterozygosity had a more modest effect (2.4-fold increase, $P = 2 \times 10^{-4}$). 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.90-9 106 deletion 5′-TCTCTGCAG TGAGATTTACTGCCACG-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%; $P = 6 \times 10^{-12}$) and insulin (-92.3%; $P = 1 \times 10^{-4}$). 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[18,19].

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)

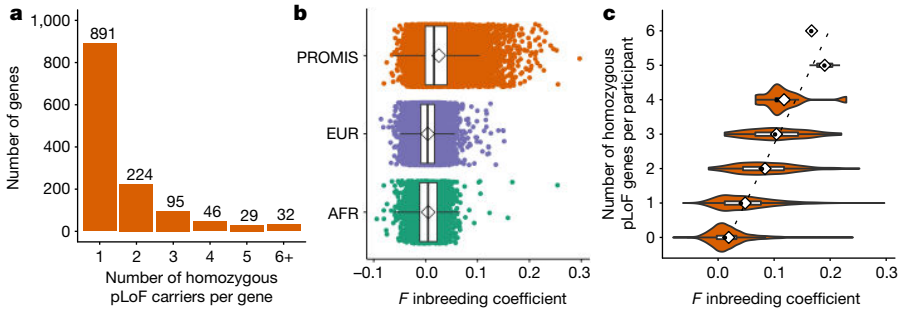


Figure 1 | Homozygous pLoF burden in PROMIS is driven by excess autozygosity. a, Most genes are observed in the homozygous pLoF state in only single individuals. **b**, The distribution of *F* inbreeding coefficient of PROMIS participants is compared to those of outbred samples of African (AFR) and European (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**).

and with the requirement that there was also evidence for association in heterozygotes (see Methods). This procedure highlighted neureglin 4 (*NRG4*), a member of the epidermal growth factor family extracellular ligands that is highly expressed in brown fat, particularly during adipocyte differentiation[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[21]. 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 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 signalling 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; *Slc9a3r1*$^{-/-}$ mice display hyperphosphaturia and disrupted protein-kinase-A-dependent cAMP-mediated phosphorylation[22]. Humans carrying rare missense mutations in *SLC9A3R1* have nephrolithiasis, osteoporosis, and hypophosphatemia[23].

Apolipoprotein C3 (apoC3, encoded by *APOC3*) is a major protein component of chylomicrons, very low-density lipoprotein cholesterol, and high-density lipoprotein cholesterol[24]. We and others recently reported that heterozygous *APOC3* pLoF mutations lower plasma triglycerides and reduce risk for coronary heart disease[4,5,25]; there is now substantial interest in *APOC3* as a therapeutic target[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 dl^{-1}, $P = 3 \times 10^{-8}$); 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 dl^{-1} over 6 h

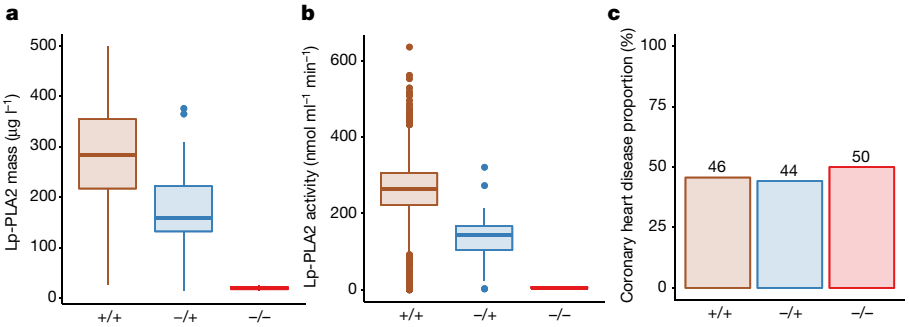


Figure 2 | Carriers of *PLA2G7* splice mutation have diminished Lp-PLA2 mass and activity but similar risk for coronary heart disease when compared to non-carriers. a, **b**, Carriage of a splice-site mutation, c.663 + 1G>A, in *PLA2G7* leads to a dose-dependent reduction of both lipoprotein-associated phospholipase A2 (Lp-PLA2) mass ($P = 6 \times 10^{-5}$) 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**).

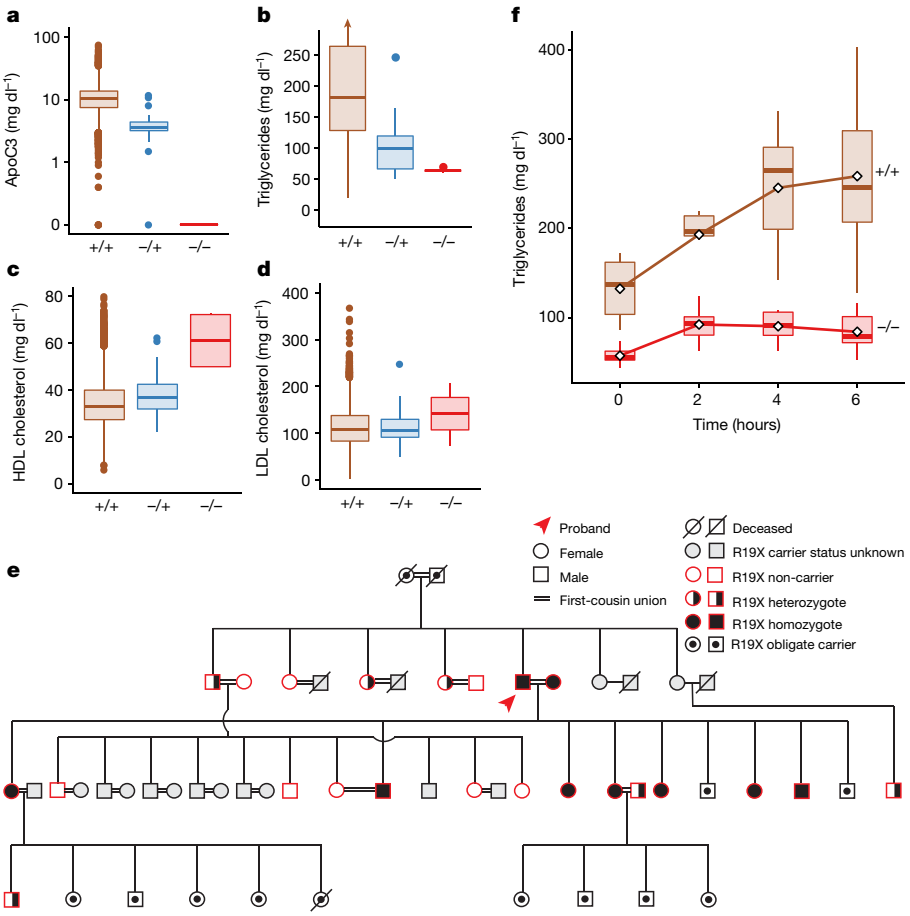


Figure 3 | *APOC3* pLoF homozygotes have diminished fasting triglycerides and blunted post-prandial lipaemia. a–d, *APOC3* pLoF genotype status, apolipoprotein C3, triglycerides, HDL cholesterol and LDL cholesterol distributions among all sequenced participants. Apolipoprotein C3 concentration is displayed 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**).

versus 1,267.7 mg dl^{-1} over 6 h; $P = 1 \times 10^{-4}$) (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[25].

Targeted gene disruption in model organisms followed by phenotypic analysis has been a fruitful approach in understanding gene function[29]; 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[6–8]. We found 56 variants

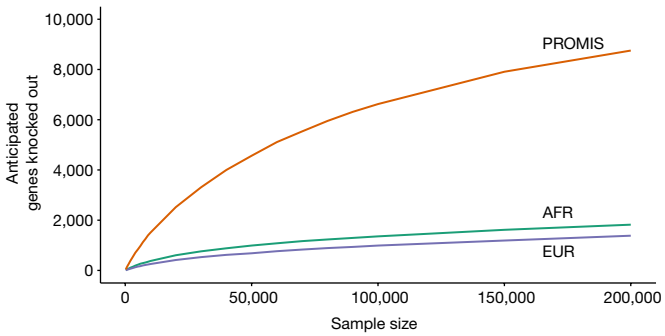


Figure 4 | Simulations anticipate many more homozygous pLoF genes in the PROMIS cohort. Number of unique homozygous pLoF genes anticipated with increasing sample sizes sequenced in PROMIS compared with similar African (AFR) and European (EUR) sample sizes. Estimates derived using observed allele frequencies 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 phenotypes 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[6,30]; (2) deep-coverage sequencing of the protein-coding regions of the genome[7]; (3) availability of a broad array of biochemical as well as clinical phenotypes across the population; (4) ability to re-contact human knockouts and their family members; (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 display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online 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 behavior and health in consanguineous marriages. *Science* **252,** 789–794 (1991).
3. 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 (2009).
4. Crosby, J. *et al.* Loss-of-function mutations in APOC3, triglycerides, and coronary disease. *N. Engl. J. Med.* **371,** 22–31 (2014).
5. Jørgensen, A. B., Frikke-Schmidt, R., Nordestgaard, B. G. & Tybjærg-Hansen, A. Loss-of-function mutations in APOC3 and risk of ischemic vascular disease. *N. Engl. J. Med.* **371,** 32–41 (2014).
6. Narasimhan, V. M. *et al.* Health and population effects of rare gene knockouts in adult humans with related parents. *Science* **352,** 474–477 (2016).
7. Sulem, P. *et al.* Identification of a large set of rare complete human knockouts. *Nat. Genet.* **47,** 448–452 (2015).
8. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).
9. Di Angelantonio, E. *et al.* Lipid-related markers and cardiovascular disease prediction. *J. Am. Med. Assoc.* **307,** 2499–2506 (2012).
10. Gregson, J. M. *et al.* Genetic invalidation of Lp-Pla2 as a therapeutic target: large-scale study of five functional Lp-Pla2-lowering alleles. *Eur. J. Prev. Cardiol.* (2016).
11. 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 (2015).
12. White, H. D. *et al.* Darapladib for preventing ischemic events in stable coronary heart disease. *N. Engl. J. Med.* **370,** 1702–1711 (2014).
13. 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. *J. Am. Med. Assoc.* **312,** 1006–1015 (2014).
14. 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 (2003).
15. 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 (1990).
16. 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).
17. 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 (2008).
18. 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 (2006).
19. 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 (2010).
20. Schneider, M. R. & Wolf, E. The epidermal growth factor receptor ligands at a glance. *J. Cell. Physiol.* **218,** 460–466 (2009).
21. 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 (2014).
22. 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 (2007).
23. Karim, Z. *et al.* NHERF1 mutations and responsiveness of renal parathyroid hormone. *N. Engl. J. Med.* **359,** 1128–1135 (2008).
24. Huff, M. W. & Hegele, R. A. Apolipoprotein C-III: going back to the future for a lipid drug target. *Circ. Res.* **112,** 1405–1408 (2013).
25. Pollin, T. I. *et al.* A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent cardioprotection. *Science* **322,** 1702–1705 (2008).
26. Gaudet, D. *et al.* Antisense inhibition of apolipoprotein C-III in patients with hypertriglyceridemia. *N. Engl. J. Med.* **373,** 438–447 (2015).
27. Gaudet, D. *et al.* Targeting APOC3 in the familial chylomicronemia syndrome. *N. Engl. J. Med.* **371,** 2200–2206 (2014).
28. 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 (2013).
29. Brown, S. D. & Moore, M. W. Towards an encyclopaedia of mammalian gene function: the International Mouse Phenotyping Consortium. *Dis. Model. Mech.* **5,** 289–292 (2012).
30. 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.-H.W. 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 (R01HL107816), the Donovan Family Foundation, and Fondation Leducq. Exome sequencing was supported by a grant from the NHGRI (5U54HG003067-11) 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 NIHR 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 NIHR Cambridge Biomedical Research Centre. In recognition for PROMIS fieldwork and 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. Ahmed, M. W. Khan, M. R. Khan, A. Ghafoor, M. Alam, R. Ahmed, M. I. Javed, A. Ghaffar, T. B. Mirza, M. Shahid, J. Furqan, M. I. Abbasi, T. Abbas, R. Zulfiqar, M. Wajid, I. Ali, M. Ikhlaq, D. Sheikh, M. Imran, M. Walker, N. Sarwar, 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.F., J.D., A.R., M.Z., M.S., A.I., S.A., F.Ma., M.I., S.A., K.T., N.H.M., K.S.Z., N.Q., M.I., S.Z.R., F.Me., 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.M.A., K.J.K., 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. 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. 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).

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 autoanalysers. 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 (apoA1, apoA2, apoB, apoC3, 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; 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, choloride, phosphate, sex-hormone binding globulin were made 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-monocyte colony stimulating factor, granulocyte colony stimulating factor, interferon-γ, IL-1β, IL-1 receptor, IL-2, IL-3, IL-4, IL-5, IL-6, IL-7, IL-8, IL-10, IL-18, IL-p40, IL-p70, IL-15, IL-17, IL-23, macrophage inflammatory protein 1α, macrophage inflammatory protein 1β, malondialdehyde-modified LDL, 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, 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[31]. 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 $<10^{-7}$ 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 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[32].

Laboratory methods for exome sequencing. *Exome sequencing.* Exome sequencing was performed at the Broad Institute. Sequencing and exome capture methods have been previously described[33,34]. 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 2D-barcoded 0.75 ml Matrix tubes at −20 °C in the SmaRTStore (RTShester, 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[35], 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 adaptor 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 adaptor sequence to facilitate downstream pooling. With the exception of the palindromic forked adapters, the reagents used for end repair, A-base addition, adaptor 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[35], 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 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 2 nM 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 analysed using RTA v.1.12.4.2. Each pool of whole-exome libraries was run on paired 76-bp runs, and an 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)[36]. 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)[37–39]. 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 tool from the GATK to generate a gVCF (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 of from historical values was noted.

Data processing and quality control of exome sequencing. *Variant annotation.* Variants were annotated using Variant Effect Predictor[40] and the LOFTEE[41] 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 examines the proportion of non-reference bases at reference sites, and excluded samples with high estimated contamination (FREEMIX scores >0.2)[42]. After removing monozygotic twins or duplicate samples using the KING software[43], we removed outlier samples with too many or too few SNPs (>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 SNPs 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 (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 described[44]. 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 inflation of association, 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 6.

Methods for inbreeding analyses. *Array-derived runs of homozygosity.* Analyses were conducted in PLINK[32] 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 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[45]. 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 R^2, 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 descent[46]. 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) ($P = 3.6 \times 10^{-37}$), East Asian (CHB, Han Chinese in Beijing, China; JPT, Japanese in Tokyo, Japan; CHD,= Chinese in Denver, Colorado) ($P = 5.4 \times 10^{-48}$) and African ancestries (YRI, Yoruban in Ibadan, Nigeria; MKK, Maasai in Kinyawa, Kenya) ($P = 1.3 \times 10^{-40}$), 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 their exome sequenced at the Broad Institute (Cambridge, Massachusetts) by the Myocardial Infarction Genetics consortium[34]. We extracted approximately 5,000 high-quality polymorphic SNPs in linkage equilibrium present on both target intervals that passed variant quality control metrics based on HapMap3 data[47]. Using PLINK, we estimated the coefficient of inbreeding separately within each ethnicity group[32]. The coefficient of inbreeding was estimated as the observed degree of homozygosity compared with the anticipated homozygosity derived from an estimated common ancestor[48]. 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 $(1 - F) \times \mathrm{CMAF}^2 + F \times \mathrm{CMAF}$, 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 ($X \sim (n, \mathrm{CMAF})$), 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 Consortium[49,50]. In brief, 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 χ^2 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 χ^2 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; $P < 1 \times 10^{-10}$). 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%)[51].

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 have been associated with embryonic or perinatal lethality as homozygous pLoF in mice[52]. Furthermore, 56 genes predicted to be essential using mouse/human conservation data[53] are tolerated as homozygous pLoF in Pakistani adults. In fact, nine genes are in both datasets and are also modelled as loss-of-function intolerant[50]. One such gene, *EP400* (also known as *p400*), influences cell cycle regulation via chromatin remodelling[54] and is critical for maintaining the identity of murine embryonic stem cells[55] 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 (at *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 pairings. 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 analyses[56]. For lipoprotein-related traits, the use of lipid-lowering therapy was used as a covariate. For glycemic biomarkers, only nondiabetics were used in the analysis. The *P* threshold for statistical significance was $0.05 / 18{,}959 = 3 \times 10^{-6}$.

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 ($|Z\ \text{score}| > 5$) 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.

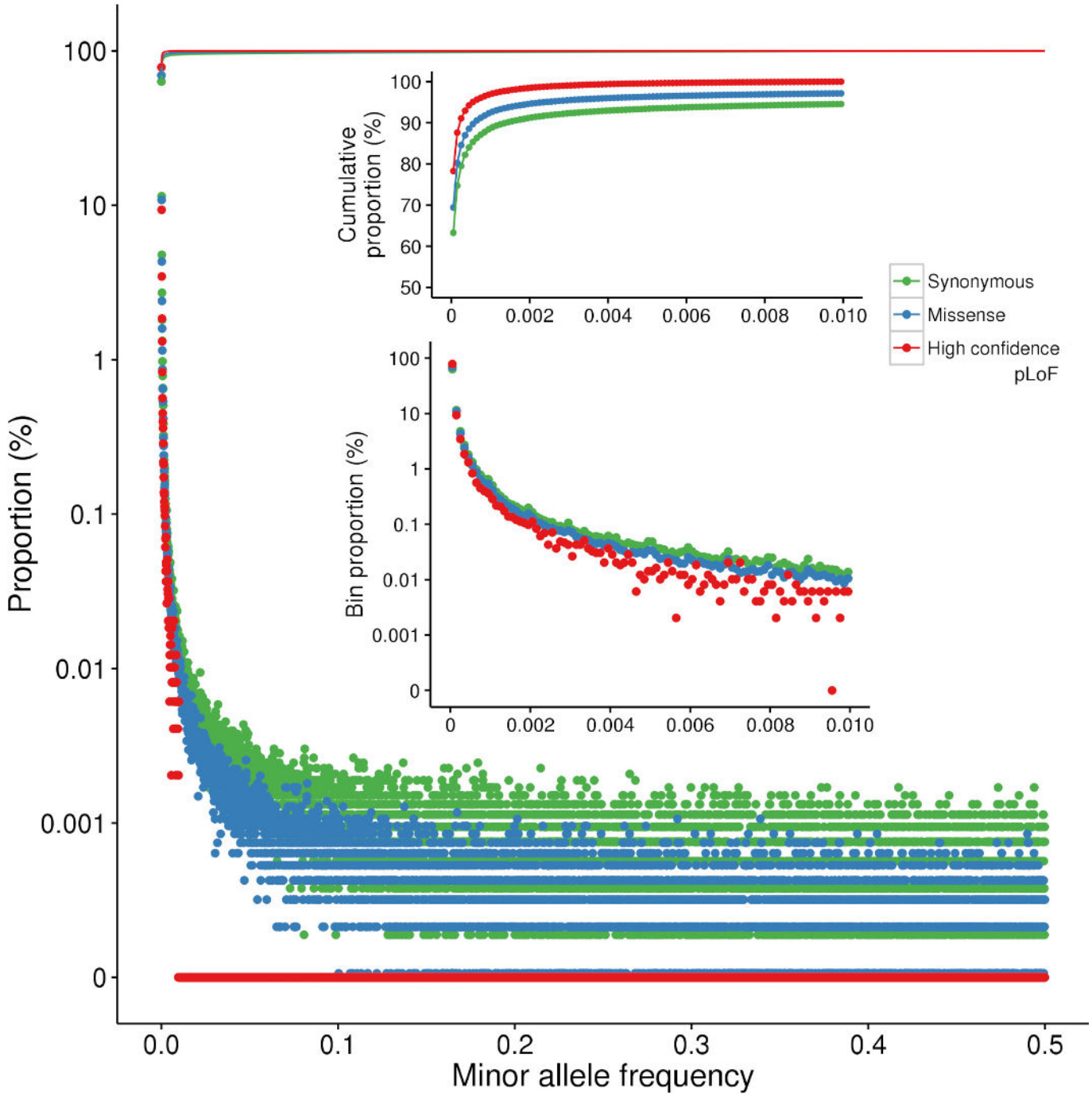
Recessive model association discovery for proteomics. Among the 84 participants with proteomic analyses of 1,310 protein analytes, 9 genes were observed in the homozygous pLoF state at least twice. We log transformed each analyte and associated with homozygous pLoF genotype status, adjusting for proteomic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

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 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, Paksitan). DNA was isolated from whole blood using a reference phenol–chloroform protocol[57]. 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′-CTCCTTCTGGCAGACCCAGCTAAGG-3′, reverse primer 5′-CCTAGGACTGCTCCGGGGAGAAAG-3′. PCR products were purified with Exo-SAP-IT (Affymetrix) and sequenced using Sanger sequencing using the same primers.

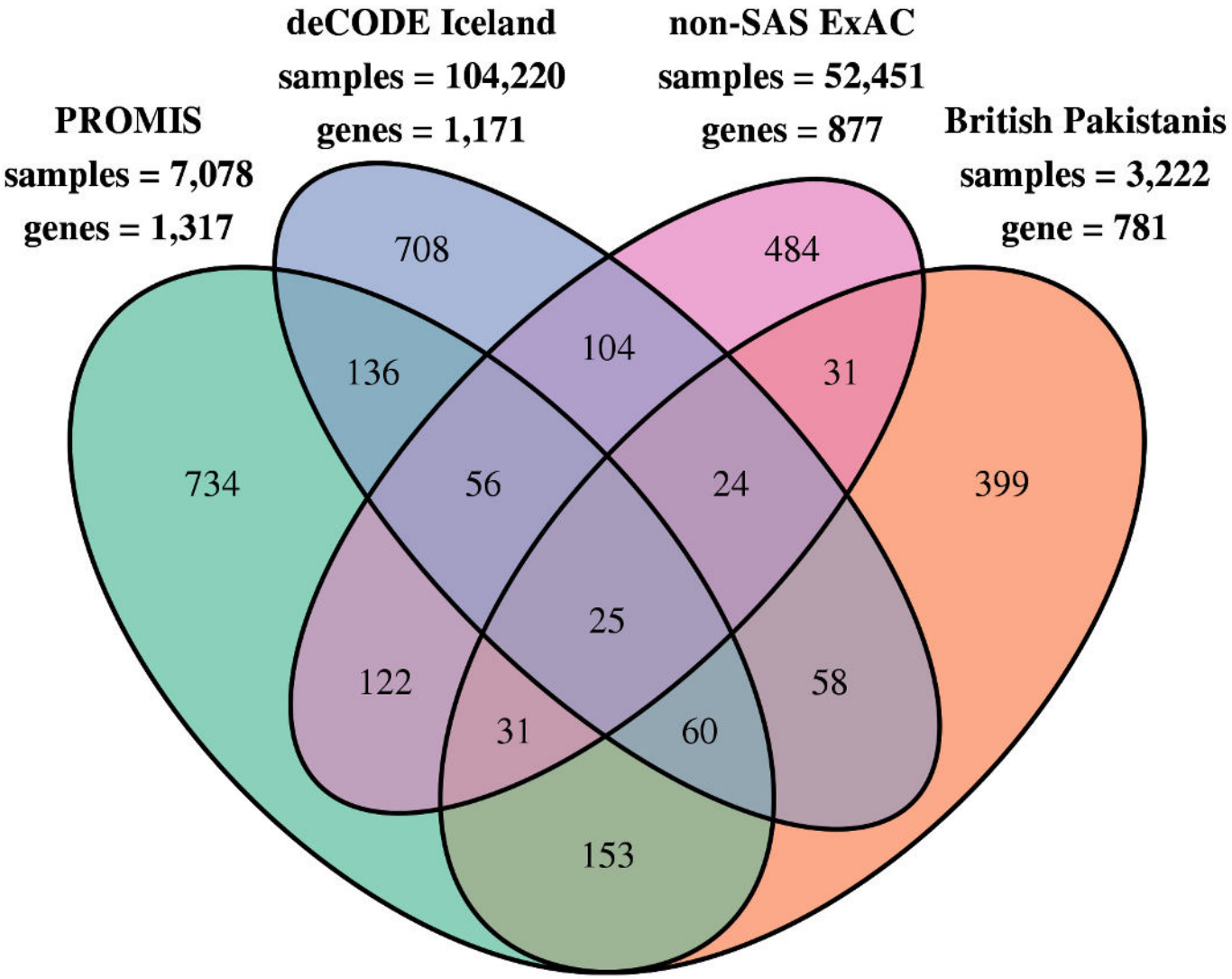
Oral fat tolerance test. Six non-carriers and seven homozygous carriers also participated in an oral fat tolerance test. Participants fasted overnight and then blood was drawn for measurement of baseline fasted lipids. Following this, participants were administered an oral load of heavy cream (50 g fat per square meter of body surface area as calculated by the method 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 previously[59]. 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, with 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.

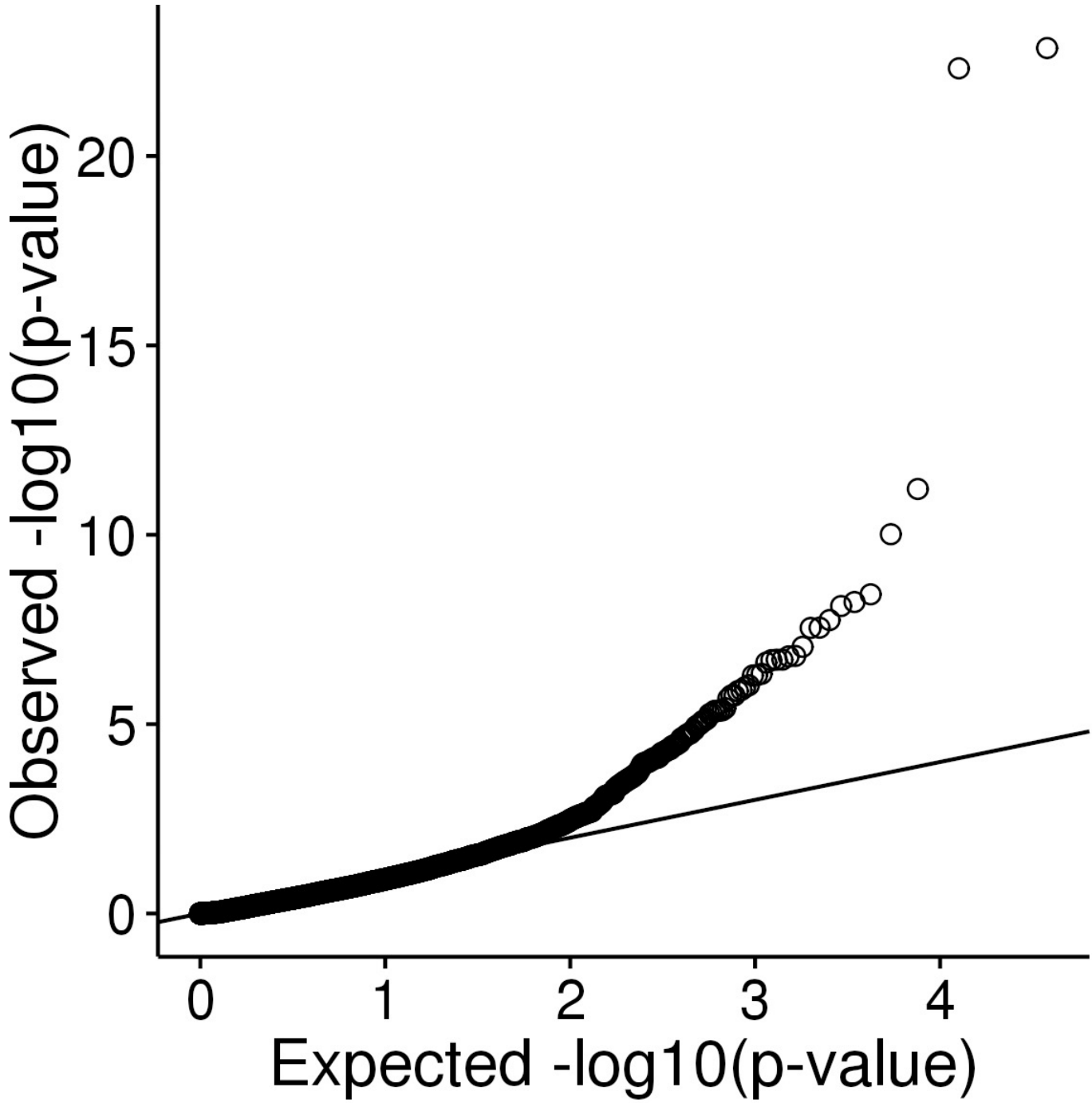
31. Kooner, J. S. *et al.* Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nat. Genet.* **43,** 984–989 (2011).
32. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
33. Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337,** 64–69 (2012).
34. Do, R. *et al.* Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. *Nature* **518,** 102–106 (2015).
35. Fisher, S. *et al.* A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biol.* **12,** R1 (2011).
36. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
37. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).
38. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–498 (2011).
39. 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* http://dx.doi.org/10.1002/0471250953.bi1110s43 (2013).
40. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26,** 2069–2070 (2010).
41. Karczewski, K. J. *Loftee (Loss-of-Function Transcript Effect Estimator),* https://github.com/konradjk/loftee (2015).
42. 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 (2012).
43. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26,** 2867–2873 (2010).
44. Gold, L. *et al.* Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoS One* **5,** e15004 (2010).
45. Hunter-Zinck, H. *et al.* Population genetic structure of the people of Qatar. *Am. J. Hum. Genet.* **87,** 17–25 (2010).
46. 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).
47. Purcell, S. M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506,** 185–190 (2014).
48. Wright, S. Coefficients of Inbreeding and Relationship. *Am. Nat.* **56,** 330–338 (1922).
49. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515,** 209–215 (2014).
50. Samocha, K. E. *et al.* A framework for the interpretation of *de novo* mutation in human disease. *Nat. Genet.* **46,** 944–950 (2014).
51. Wang, T. *et al.* Identification and characterization of essential genes in the human genome. *Science* **350,** 1096–1101 (2015).
52. Eppig, J. T., Blake, J. A., Bult, C. J., Kadin, J. A. & Richardson, J. E. The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. *Nucleic Acids Res.* **43,** D726–D736 (2015).
53. Georgi, B., Voight, B. F. & Bućan, M. From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. *PLoS Genet.* **9,** e1003484 (2013).
54. Fuchs, M. *et al.* The p400 complex is an essential E1A transformation target. *Cell* **106,** 297–307 (2001).
55. 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 (2008).
56. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38,** 904–909 (2006).
57. Sambrook, J. & Russell, D. W. Purification of nucleic acids by extraction with phenol:chloroform. *CSH Protoc.* http://dx.doi.org/10.1101/pdb.prot4455 (2006).
58. Mosteller, R. D. Simplified calculation of body-surface area. *N. Engl. J. Med.* **317,** 1098 (1987).
59. Maraki, M. *et al.* Validity of abbreviated oral fat tolerance tests for assessing postprandial lipemia. *Clin. Nutr.* **30,** 852–857 (2011).



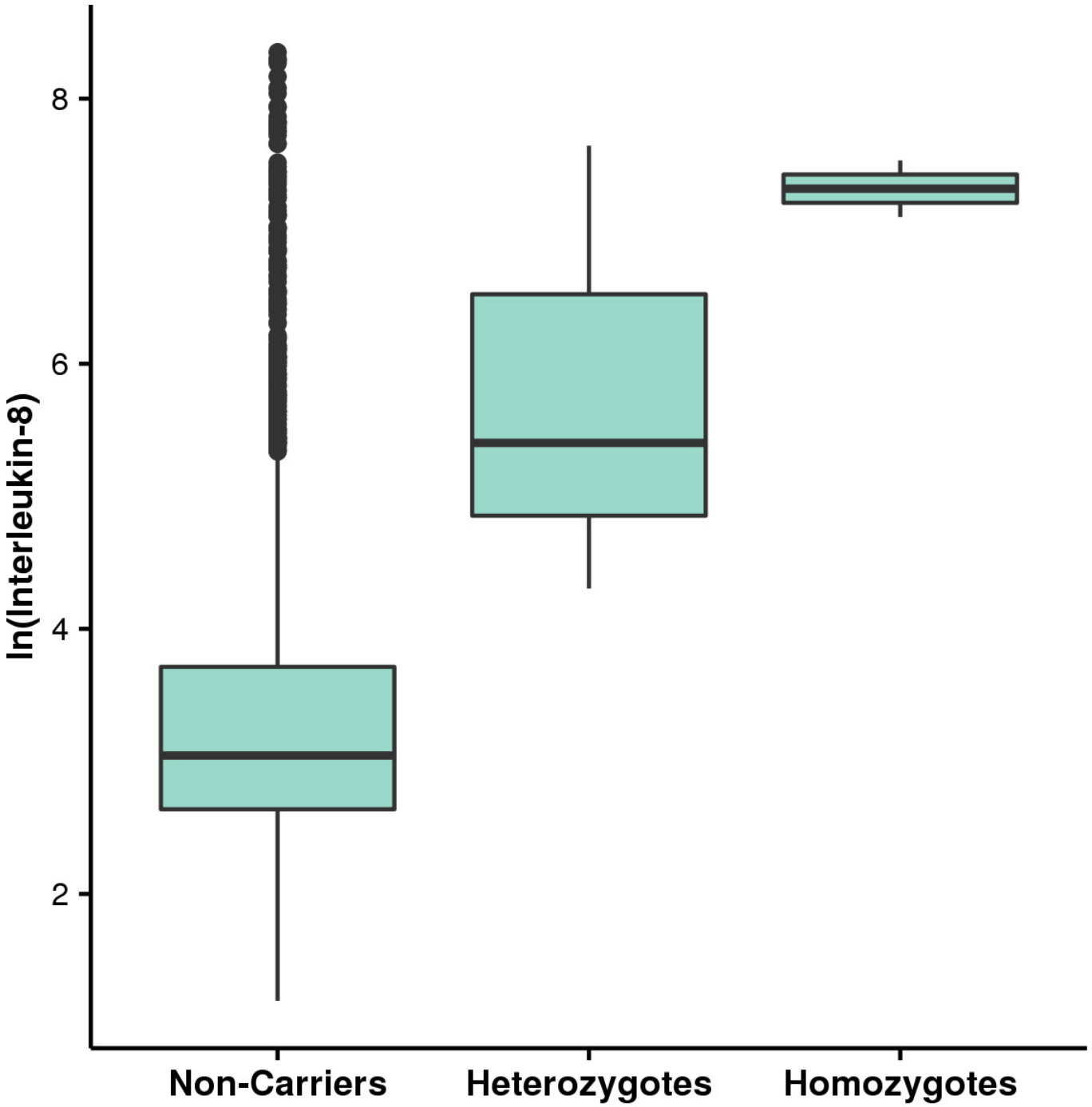
Extended Data Figure 1 | pLoF mutations are typically seen in very few individuals. The site-frequency spectrum of synonymous, missense, and high-confidence pLoF mutations is represented. Points represent the proportion of variants within a 1×10^{-4} minor allele frequency bin for each variant category. 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.



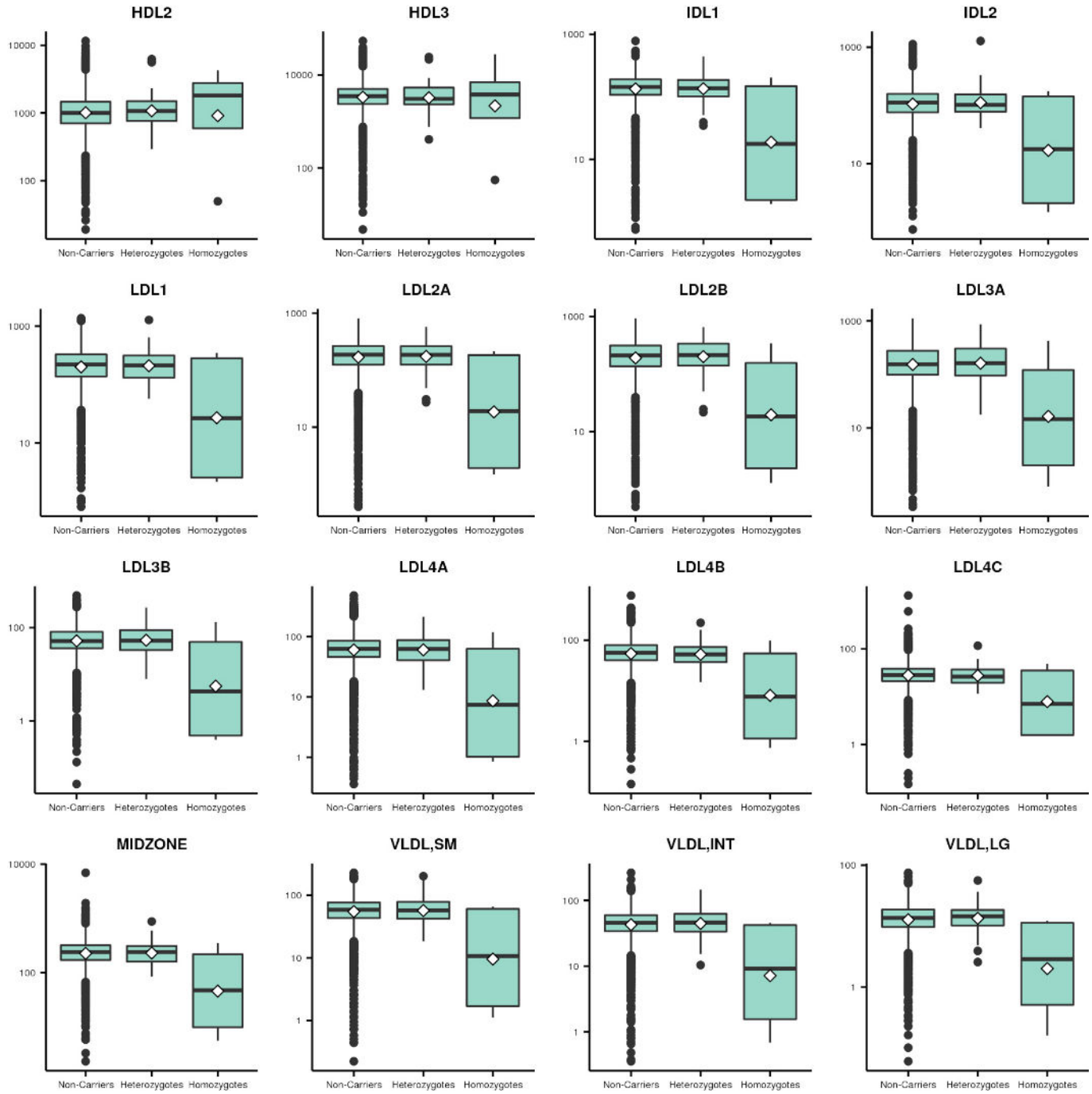
Extended Data Figure 2 | Intersection of homozygous pLoF genes between PROMIS and other cohorts. We compared the counts and overlap of unique homozygous pLoF genes in PROMIS with other exome sequenced cohorts.



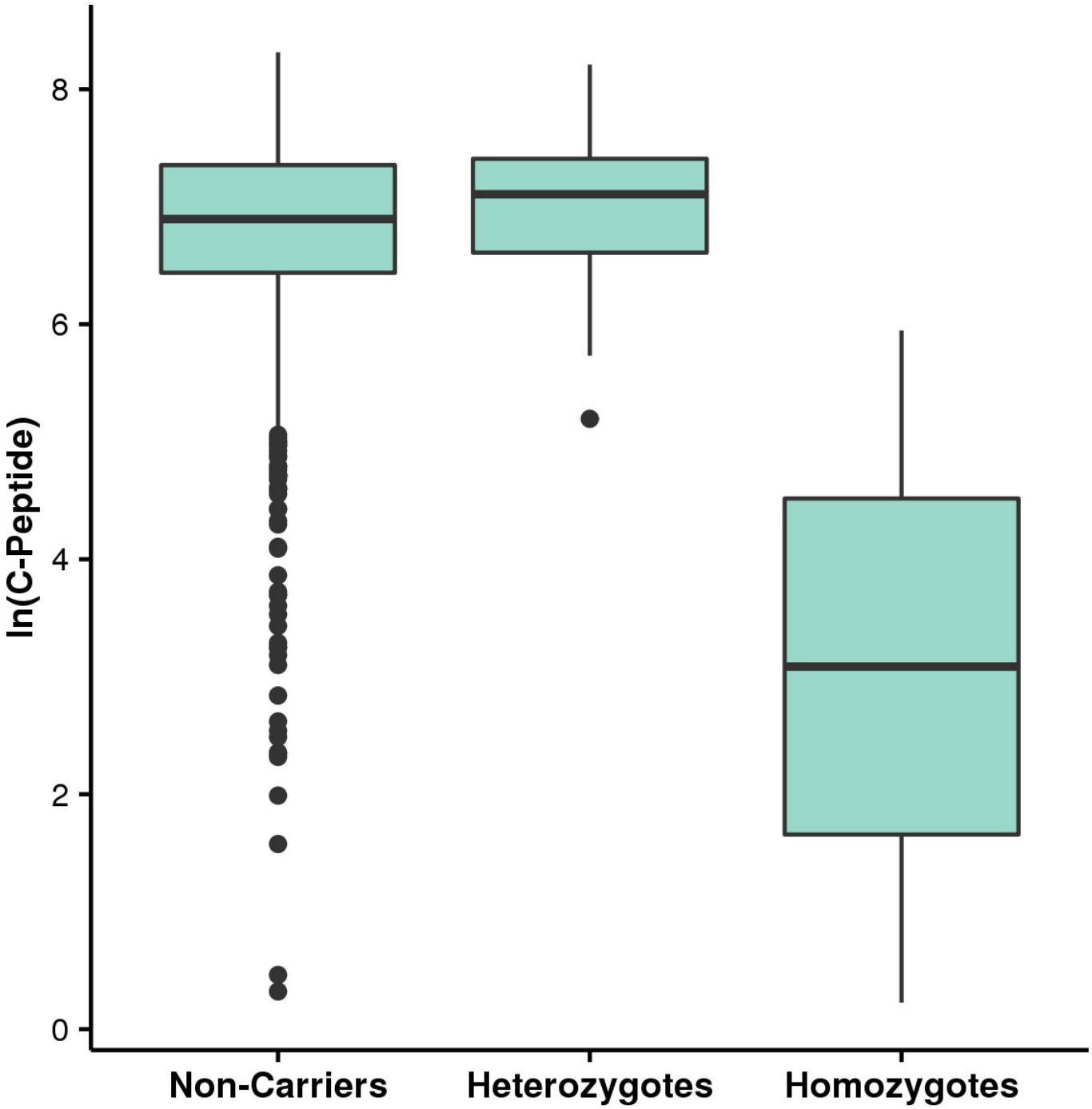
Extended Data Figure 3 | QQ-plot of recessive model pLoF association analysis across phenotypes. Analyses to determine whether homozygous pLoF carrier status was associated with traits was performed where there were at least two homozygous pLoF carriers phenotyped per trait. The observed versus the expected results from 15,263 associations are displayed here demonstrating an excess of associations beyond a Bonferroni threshold.



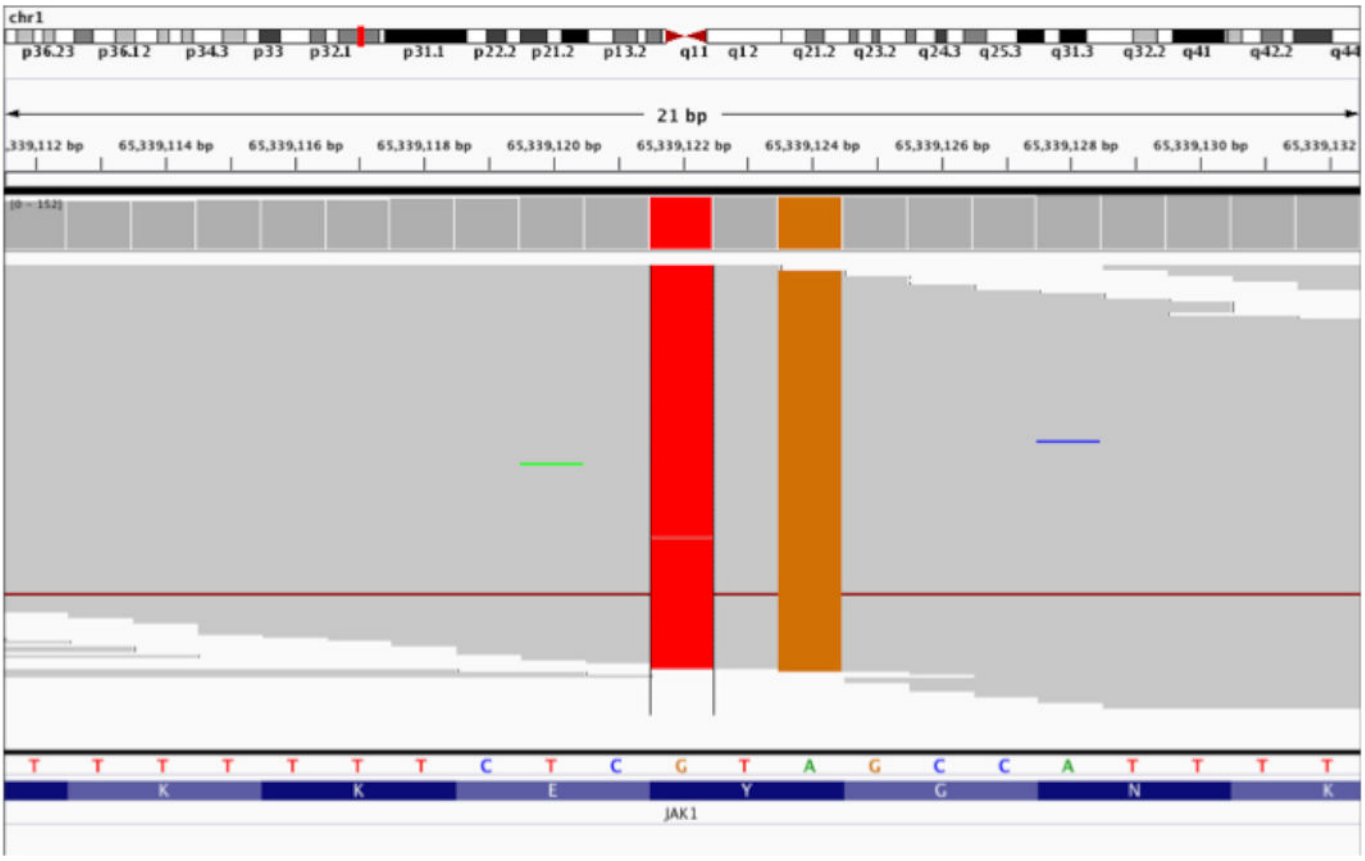
Extended Data Figure 4 | Carriers of pLoF alleles in *CYP2F1* have increased IL-8 concentrations. Participants who had pLoF mutations in the *CYP2F1* gene had higher concentrations of IL-8, whereas heterozygotes had a more modest effect when compared to the rest of the cohort of non-carriers. IL-8 concentration is natural log transformed. Bars represent 1.5× interquartile range beyond the 25th and 75th percentiles.

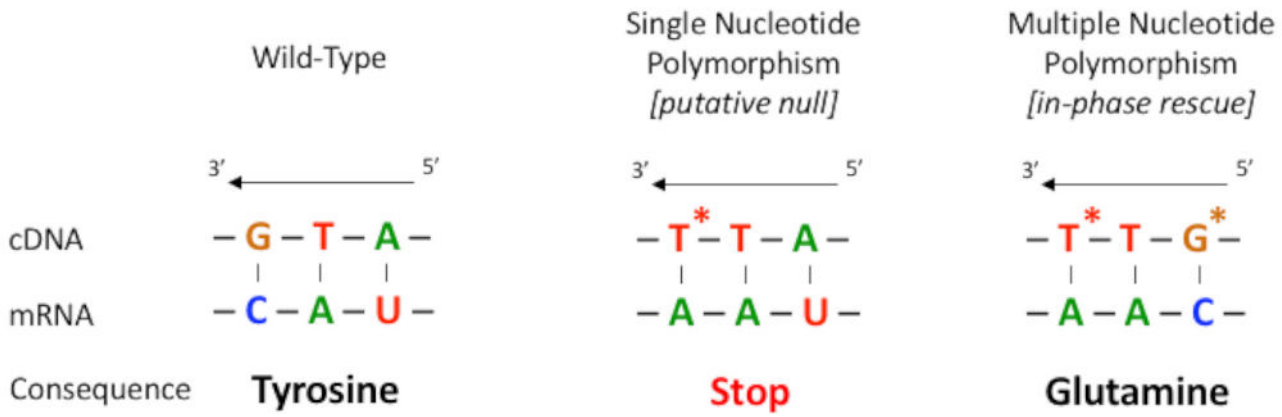


Extended Data Figure 5 | Carriers of pLoF alleles in *TREH* have decreased concentrations of several lipoprotein subfractions. Participants who had pLoF mutations in the *TREH* gene had lower concentrations of several lipoprotein subfractions. Bars represent 1.5× interquartile range beyond the 25th and 75th percentiles.

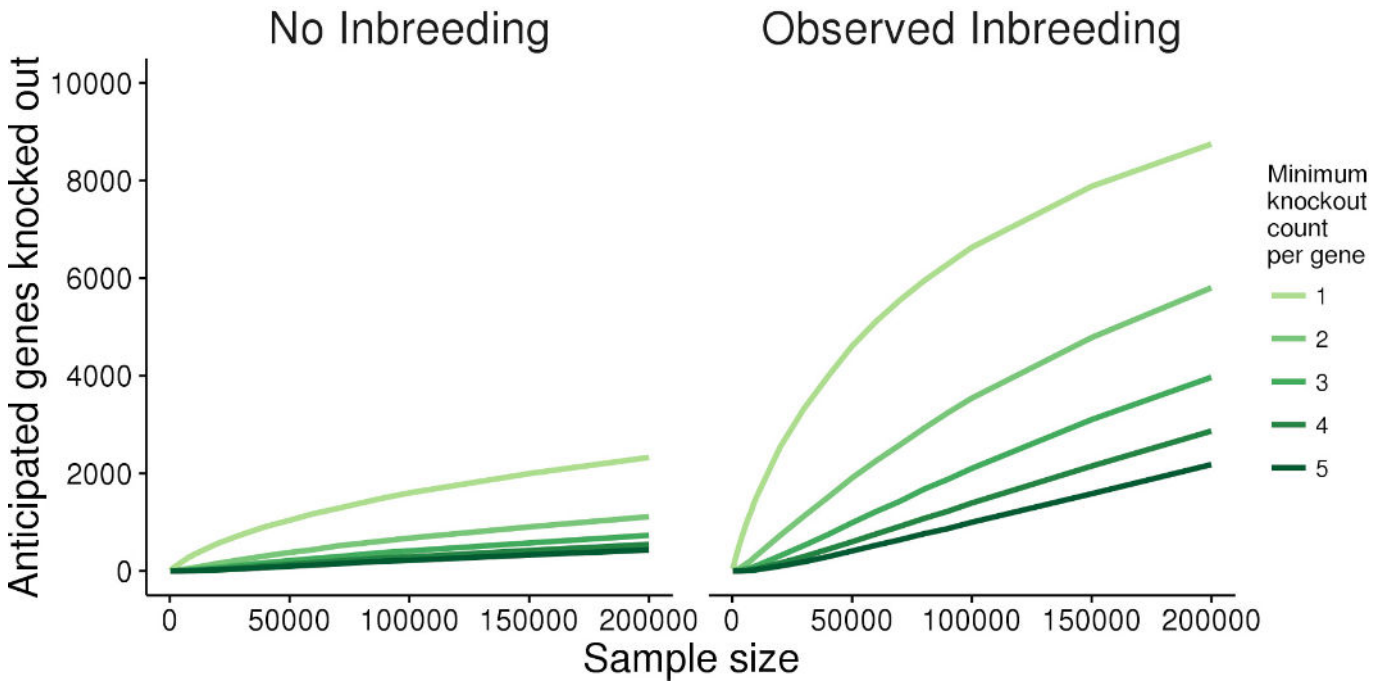


Extended Data Figure 6 | Nondiabetic homozygous pLoF carriers for *A3GALT2* have diminished insulin C-peptide concentrations. Among nondiabetics, those who were homozygous pLoF for *A3GALT2* had substantially lower fasting insulin C-peptide concentrations. This observation was not evident in nondiabetic heterozygous pLoF *A3GALT2* participants. Insulin C-peptide is natural log transformed. Bars represent 1.5× interquartile range beyond the 25th and 75th percentiles.

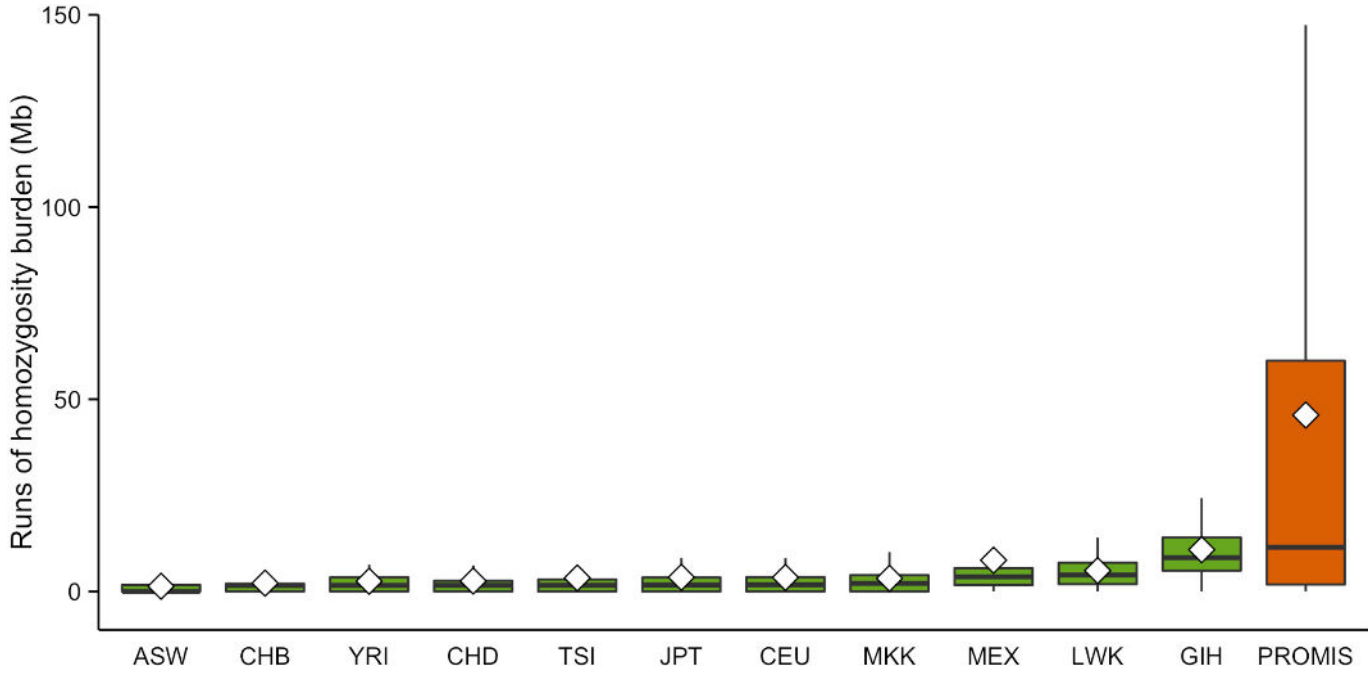




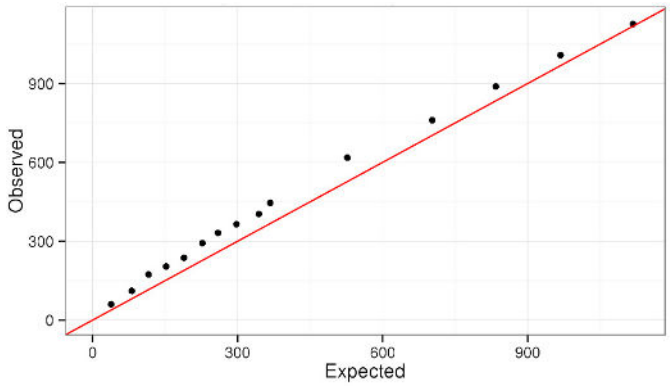
Extended Data Figure 7 | Example of a second polymorphism in-phase which rescues a putative protein-truncating mutation. Short-reads that align to genomic positions 65,339,112 to 65,339,132 on chromosome 1 are displayed for one individual with a putative homozygous pLoF genotype in this region. The SNP 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.

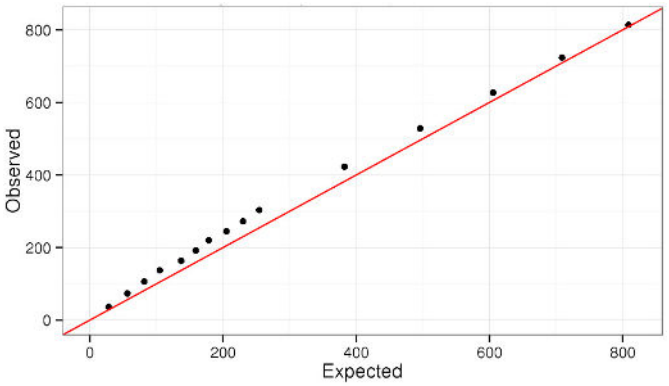


Extended Data Figure 8 | Anticipated number of genes knocked out with increasing sample sizes by minimum knockout count. We simulate the number of genes expected to be knocked out by minimum knockout count per gene at increasing sample sizes. We perform this simulation with and without the observed inbreeding.



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.