Proteome-by-phenome Mendelian Randomisation detects 38 proteins with causal roles in human diseases and traits

ANDREW D. BRETHERICK[1*], ORIOL CANELA-XANDRI[1,2], PETER K. JOSHI[3], DAVID W. CLARK[3], KONRAD RAWLIK[2], THIBAUD S. BOUTIN[1], YANNI ZENG[1,4,5,6], CARMEN AMADOR[1], PAU NAVARRO[1], IGOR RUDAN[3], ALAN F. WRIGHT[1], HARRY CAMPBELL[3], VERONIQUE VITART[1], CAROLINE HAYWARD[1], JAMES F. WILSON[1,3], ALBERT TENESA[1,2], CHRIS P. PONTING[1], J. KENNETH BAILLIE[2], AND CHRIS HALEY[1,2*]

[1] *MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, EH4 2XU, Scotland, UK.*

[2] *The Roslin Institute, University of Edinburgh, Easter Bush, EH25 9RG, Scotland, UK.*

[3] *Centre for Global Health Research, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Teviot Place, Edinburgh, EH8 9AG, Scotland, UK.*

[4] *Faculty of Forensic Medicine, Zhongshan School of Medicine, Sun Yat-Sen University, 74 Zhongshan 2nd Road, Guangzhou 510080, China.*

[5] *Guangdong Province Translational Forensic Medicine Engineering Technology Research Center, Zhongshan School of Medicine, Sun Yat-Sen University, 74 Zhongshan 2nd Road, Guangzhou 510080, China.*

[6] *Guangdong Province Key Laboratory of Brain Function and Disease, Zhongshan School of Medicine, Sun Yat-Sen University, 74 Zhongshan 2nd Road, Guangzhou 510080, China.*

** Corresponding Author*

Abstract. Target identification remains a crucial challenge in drug development. To enable unbiased detection of proteins and pathways that have a causal role in disease pathogenesis or progression, we propose proteome-by-phenome Mendelian Randomisation (P²MR). We first detected genetic variants associated with plasma concentration of 249 proteins. We then used 64 replicated variants in two-sample Mendelian Randomisation to quantify evidence of a causal role for each protein across 846 phenotypes: this yielded 509 robust protein-outcome links. P²MR provides substantial promise for drug target prioritisation. We provide confirmatory evidence for a causal role for the proteins encoded at multiple cardiovascular disease risk loci (*FGF5*, *IL6R*, *LPL, LTA*), and discovered that intestinal fatty acid binding protein (*FABP2*) contributes to disease pathogenesis. Additionally, we find and replicate evidence for a causal role of tyrosine-protein phosphatase non-receptor type substrate 1 (SHPS1; *SIRPA*) in schizophrenia. Our results provide specific prediction of the effects of changes of plasma protein concentration on complex phenotypes in humans.

An initial goal of drug development is the identification of targets – in most cases, proteins – whose interaction with a drug ameliorates the development, progression, or symptoms of disease. After some success, the rate of discovery of new targets has not accelerated despite substantially increased investment (Munos, 2009). A large proportion of drugs fail at the last stages of development – clinical trials – because their targets do not alter whole-organism phenotypes as expected from pre-clinical research (Arrowsmith, 2011).

Preclinical science is engaging with increasingly complex systems in which prediction of the effects of an intervention is ever more difficult (Civelek & Lusis, 2014). The ability to cut through complexity to distinguish factors that modulate whole-organism phenotypes is a major advantage of genetic (Baillie, 2014) and functional genomic (Baillie et al., 2018) approaches to drug development. Nevertheless, genetic associations with disease are not immediately interpretable (MacArthur et al., 2017): most disease-associated variants fail to alter protein-coding sequence, but instead alter protein levels via often poorly understood molecular mechanisms.

A subset of disease states have been studied with adequately-powered genome-wide association (GWA) studies (Finan et al., 2017). From these, persuasive evidence already exists for the utility of using genetic and genomic techniques to inform drug development: the presence of genetic evidence in support of a protein could double the probability of success in clinical trials for drugs targeting that protein (M. R. Nelson et al., 2015). In a recent study, 12% of all targets for licenced drugs could be rediscovered using GWA studies (Finan et al., 2017). However, these GWA study approaches generally rely on measures of proximity of a disease-associated genetic variant to a protein-coding gene, and proximity alone does not imply causality.

Mendelian Randomisation (MR) uses genetic variants to provide an estimate of the effect of an exposure on an outcome, using the randomness of assignment of genotype to remove the effects of unmeasured confounding (Smith & Ebrahim, 2003). The approach is analogous to

a naturally-occurring randomised control trial. When a genetic variant predicts the abundance of a mediator, MR tests the hypothesis that this mediator plays a causal role in disease risk. This is possible because the patient or participant was effectively randomised at conception to a genetically-determined level of that mediator. Under this model, it is possible to use population level genetic information to draw causal inference from observational data. However, there are, as with any study, unverifiable assumptions with this study design: a major concern is that alternative causal pathways may link the instrumental variable (here, the DNA variant) to the phenotype (the disease outcome). In a clinical trial this would be analogous to a drug influencing a disease through a different pathway than via its reported target. In MR, addressing the risk of alternative causal pathways is of great practical importance in order to avoid pursuing drugs that target an irrelevant molecular entity, and hence that have no beneficial effect. In order to address this, we limited ourselves to using locally-acting pQTLs as instrumental variables. We believe this approach provides stronger supporting evidence for causation than relying on proximity of a disease-associated genetic variant to a gene, or using mRNA abundance as a proxy for protein abundance (Mirauta et al., 2018).

Due to recent advances in proteomic technologies, the availability of pQTL data has increased dramatically in recent years (Folkersen et al., 2017; Suhre et al., 2017; Sun et al., 2018; Yao et al., 2018). A number of these studies attempt to infer causality using MR and similar techniques. In our approach, we applied pQTL based MR in a data-driven manner across the full range of phenotypes available in GeneAtlas (UK Biobank (Canela-Xandri, Rawlik, & Tenesa, 2017)), as well as supplementing this with additional studies identified through Phenoscanner (Staley et al., 2016). We performed GWA for 249 proteins in two European cohorts, and then adopted a proteome-by-phenome Mendelian randomisation (P^2MR) approach to assess the potential causal role of 64 proteins in 846 outcomes (e.g. diseases, anthropomorphic measures, etc.). GeneAtlas results were further stratified according to their consistency with a single underlying causal variant (affecting both variation in protein concentration and outcome phenotype) or otherwise. Ultimately, of the

249 proteins, 38 were identified as causally contributing to human disease or other quantitative trait.

Results

The abundance of an individual protein may be associated with DNA variants both local and distant to its gene (termed local- and distal-pQTLs, respectively). We assayed the plasma levels of 249 proteins using high-throughput, multiplex immunoassays and then performed genome-wide association of these levels in two independent cohorts (discovery and replication) of 909 and 998 European individuals. P^2MR was applied to 54,144 exposure-outcome pairs obtained from 64 significantly (p-value $<5x10^{-8}$) associated, replicated (Bonferroni correction for multiple testing), local-pQTLs, and 778 phenotypes obtained from GeneAtlas (UK Biobank (Canela-Xandri et al., 2017)) and 68 phenotypes from 20 additional genome-wide association (meta-analysis) studies (The CARDIoGRAMplusC4D Consortium et al., 2015; R. A. Scott et al., 2017; C. P. Nelson et al., 2017; Liu et al., 2015; Schizophrenia Working Group of the Psychiatric Genomics Consortium et al., 2014; Bronson et al., 2016; Okada et al., 2014; van Rheenen et al., 2016; Hammerschlag et al., 2017; Sniekers et al., 2017; Okbay et al., 2016; Hou et al., 2016; Beaumont et al., 2018; Phelan et al., 2017; van der Harst & Verweij, 2018; Berg et al., 2016; de Moor et al., 2015; The EArly Genetics and Lifecourse Epidemiology (EAGLE) Eczema Consortium et al., 2015; M. A. Ferreira et al., 2017; Astle et al., 2016) identified through Phenoscanner (Staley et al., 2016) (Figure 1; Supplementary Table S1; Methods).

In total, we identified 509 protein-outcome links for which there is evidence of a causal role of the exposure (protein) on the outcome phenotype (trait).

pQTLs. pQTLs were highly concordant between the two cohorts (Supplementary Table S2). Of the 209 independent pQTLs identified in the discovery cohort (p-value $<5x10^{-8}$), 154 were successfully replicated (Bonferroni correction for multiple testing; consistent direction of

effect). These represented pQTLs for 82 proteins, all but two encoded by autosomal genes. Lead variants (smallest p-value within the locus; Methods) were identified at each locus, the majority (64/80; 80%) of these proteins had the lead variant of one or more pQTLs located close to the gene encoding the protein (±150kb; Figure 1) and hence were used as instrumental variables suitable for MR. In many respects, locally-acting pQTLs are ideal instrumental variables: they have large effect sizes, have highly plausible biological relationships with protein level, and provide quantitative information about (often) directly druggable protein targets. This is in contrast to distal pQTLs: the pathway through which they exert their effects is generally unknown, with no *a priori* expectation of a direct effect on a single target gene.

Outcome GWA Studies. Results linking the genetic variants and outcome traits and diseases were obtained from secondary cohorts. UK Biobank has captured a wealth of information on a large – approximately 500,000 individuals – population cohort that includes anthropometry, haematological traits, and disease outcomes. Genome-wide association of 778 phenotypes from UK Biobank has been performed and published as GeneAtlas (Canela-Xandri et al., 2017). Although the cohort is large, for many diseases the number of UK Biobank individuals affected is small, resulting in low statistical power. Consequently, we augmented these results with additional studies identified using Phenoscanner (Staley et al., 2016) (Methods).

Mendelian Randomisation. MR depends upon an assumption that the DNA variant used as an instrumental variable is robustly associated with the exposure. In our case we ensured this by using stringent discovery and replication criteria for instrument selection. By limiting ourselves to using locally-acting pQTLs as instruments, we sought to leverage *a priori* biological knowledge regarding cellular protein production to substantially increase confidence in the existence of a direct path from DNA variant to protein, and from protein to outcome. P^{2}MR yielded 271 protein-outcome pairs that were significant (false discovery rate (FDR) <0.05) in UK Biobank, and 238 significant (FDR <0.05) pairs using data from

153 Phenoscanner. Thirty-two of the 64 proteins were causally implicated for one or more
154 outcomes in UK Biobank, and 36 of 64 in the outcomes identified through Phenoscanner
155 studies. The outcomes from GeneAtlas and Phenoscanner are not mutually exclusive, and
156 some of the studies included from Phenoscanner included data from UK Biobank, however,
157 overall, 38 of the 64 proteins (60%) were implicated in at least one outcome (Supplementary
158 Tables S3 and S4).
159
160 Proteins were implicated in diseases ranging from schizophrenia to cardiovascular disease
161 (Figure 2, Supplementary Tables S3 and S4). We applied a method, HEIDI (Zhu et al., 2016),
162 which explicitly accounts for the linkage disequilibrium (LD) structure of the locus to assess
163 the heterogeneity of MR effect estimates between the lead variant (the primary instrument)
164 and those of linked variants. HEIDI tests the hypothesis that the observed MR results are
165 caused by two distinct causal variants. Of the UK Biobank causal inferences, 77 survived the
166 HEIDI heterogeneity test (p-value >0.05). Therefore, these 77 proteins have (a) high-quality
167 evidence of association to a DNA variant which provides congruent predictions for both
168 plasma protein levels and disease risk / outcome phenotype, and (b) because of the physical
169 proximity to the SNP to the coding-sequence of the gene for the protein, and non-significant
170 HEIDI result, a low risk of pleiotropy (Supplementary Tables S3). These pairs thus provide
171 the most robust evidence that the level of the protein directly alters disease risk / outcome
172 phenotype.
173
174 However, all 509 causal inferences (271 from GeneAtlas (Canela-Xandri et al., 2017) and 238
175 from studies identified through Phenoscanner (Staley et al., 2016); Figures 2, S1, S2, S3, S4,
176 and Tables S3 and S4), even those consistent with heterogeneity, remain potential high
177 quality drug targets. This is because the HEIDI heterogeneity test (Figure 1) is susceptible to
178 type I errors in this context, as it does not account for multiple causal variants in a locus. In
179 addition, we apply HEIDI in a conservative manner: as a significant HEIDI test implies
180 heterogeneity, we did not apply a multiple testing correction. If a Bonferroni correction (271

tests) were to be applied to the HEIDI p-value, 180 of the protein-outcome pairs are not significantly heterogeneous.

For some of these inferences, genetic evidence of an association between a protein and phenotype has been proposed based on physical proximity of the genes to GWA intervals. For nearly two-thirds (62%; 318/509) however, significant (FDR <0.05) MR association between protein and outcome was not matched by significant (p-value $<5x10^{-8}$) association of the DNA variant to outcome. This suggests that P^2MR has a greater potential to link protein product and phenotype than naïve genome-wide association.

Our results draw causal inference between protein concentration and disease, for example, IL4R and asthma, IL2RA and thyroid dysfunction, and IL12B and psoriasis (Figure 2). Taking IL6R as an example, we found evidence for a causal association between plasma IL6R abundance and coronary artery disease (CAD), atopy, and rheumatoid arthritis (Figure 2, and Tables S3 and S4). We note that: 1) tocilizumab (an IL6 receptor antagonist) is in clinical use for treating rheumatoid arthritis (L. J. Scott, 2017), 2) there is prior evidence from MR demonstrating elevated levels of soluble IL6R and reduced cardiovascular disease (IL6R Genetics Consortium Emerging Risk Factors Collaboration, 2012; Interleukin-6 Receptor Mendelian Randomisation Analysis (IL6R MR) Consortium et al., 2012), and 3) the evidence of a causal link between IL6R and atopy was not well established previously. Notably however, tocilizumab has been used to treat three atopic dermatitis patients, and all patients experienced >50% improvement in disease (Navarini, French, & Hofbauer, 2011). In addition, Ullah et al. (2014) demonstrated that tocilizumab caused a reduction in Th2/Th17 response and associated airway inflammatory infiltration in a mouse model of experimental allergic asthma.

As further illustration, we take two clinically important phenotypes as case-studies: CAD risk and schizophrenia.

CAD and FABP2: P²MR identified 5 proteins as contributing to CAD pathogenesis: FABP2, FGF5, IL6R, LPL, and LTA. Of these, 4 (FGF5, LPL, IL6R, and LTA) had been implicated previously (Klarin et al., 2017; C. P. Nelson et al., 2017; Ozaki et al., 2002), whereas FABP2 had more limited evidence for its involvement.

FABP2 (intestinal fatty acid-binding protein) is causally linked by P²MR to CAD (Figure 2). A FABP2 non-synonymous mutation (Ala54Thr) had been proposed as a risk factor for CAD (Yuan, Yu, & Zeng, 2015), consistent with its P²MR candidature. However, of critical importance to its potential utility as a therapeutic target, our study validates and extends this association beyond the non-synonymous variant to protein abundance. pQTL analysis identified two lead DNA variants in close proximity (<150kb) to the *FABP2* gene. Using the SNP rs17009129, P²MR finds a causal link between FABP2 concentration and CAD ($p = 1.1x10^{-4}$; FDR <0.05; β_{MR} -0.11; se_{MR} 0.028; β_{MR} and se_{MR} units: log(OR)/standard deviation of residualised protein concentration) without significant heterogeneity (p = 0.24) which suggests shared causal genetic control. Furthermore, a second independent SNP (r^2 <0.2; rs6857105) replicates this observation (MR $p = 5.0x10^{-4}$; HEIDI p = 0.34; β_{MR} -0.17; se_{MR} 0.047). Both SNPs (rs17009129, and rs6857105) fell below genome-wide significance ($p<5x10^{-8}$) in the full meta-analysis of van der Harst (van der Harst & Verweij, 2018) on CAD; however, FABP2 was flagged as potentially relevant by DEPICT, a prioritization tool. Consequently, this is the first time, to our knowledge, that variants associate with *FABP2* concentration have been shown robustly to causally contribute to CAD pathogenesis.

Schizophrenia: By applying P²MR, we identified 3 proteins that were causally implicated in the pathogenesis of schizophrenia: (i) Tyrosine-protein phosphatase non-receptor type substrate 1 (SHPS1; *SIRPA*), (ii) Tumour necrosis factor receptor superfamily member 5 (*CD40*), and (iii) Low affinity immunoglobulin gamma Fc region receptor II-b (*FCGR2B*). The link between SHPS1 (rs4813319) and schizophrenia risk was subsequently replicated in the UK Biobank data (Methods; Table 1). The observed effect of SHSP1 on schizophrenia was not significantly heterogeneous in the results of the Schizophrenia Working Group of the

239 Psychiatric Genomics Consortium (2014) (p = 0.53). Here we investigate *SHPS1* (approved
240 symbol *SIRPA*), which encodes SHPS1, tyrosine-protein phosphatase non-receptor type
241 substrate 1 and use *SHPS1* henceforth.
242
243 Interestingly, SHPS1 is highly expressed in the brain, especially in the neuropil (a dense
244 network of axons, dendrites, and microglial cell processes) in the cerebral cortex ("SIRPA
245 available from v18.proteinatlas.org," 2018; "The Human Protein Atlas," n.d.; Thul et al.,
246 2017; Uhlén et al., 2015; Uhlen et al., 2017), and co-localises with CD47 at dendrite-axon
247 contacts (Ohnishi et al., 2005). Mouse models in which the *SHPS1* gene is disrupted exhibit
248 many nervous system abnormalities, such as reduced long term potentiation, abnormal
249 synapse morphology and abnormal excitatory postsynaptic potential (MGI: 5558020
250 ("Mouse Genome Informatics (v6.13)," 2019; Toth et al., 2013)). Other mouse and rat models
251 link CD47 to sensorimotor gating and social behaviour phenotypes (H. P. Chang, Lindberg,
252 Wang, Huang, & Lee, 1999; Huang, Wang, Tang, & Lee, 1998; Koshimizu, Takao, Matozaki,
253 Ohnishi, & Miyakawa, 2014; Ma, Kulesskaya, Võikar, & Tian, 2015; Ohnishi et al., 2010). In
254 addition, SHPS1 has been shown to mediate activity-dependent synapse maturation (Toth et
255 al., 2013) and may also have a role as a "don't eat me" signal to microglia (Brown & Neher,
256 2014). Finally, SHPS1 levels tend to be lower in the dorsolateral prefrontal cortex of
257 schizophrenia patients (Martins-de-Souza et al., 2009).
258

Table 1: Replication of significant Mendelian Randomisation (FDR <0.05) protein-to-schizophrenia links in UK Biobank.

Discovery: Mendelian randomisation p-value of protein level on schizophrenia risk, as estimated using data from the Psychiatric Genomics Consortium – PGC (Schizophrenia Working Group of the Psychiatric Genomics Consortium et al., 2014) obtained via Phenoscanner (Staley et al., 2016).

Replication: The Mendelian Randomisation p-value of protein level on schizophrenia (combined risk of 'F20-F29 Schizophrenia, schizotypal and delusional disorders' and self-reported 'schizophrenia') in UK Biobank (Methods).

MR p-values are significant (FDR <0.05) in the Discovery sample.

† indicates significance of the replication study following multiple testing correction (Bonferroni).

Gene	SNP	Discovery	Replication
SHPS1	rs4813319	1.1×10^{-3}	5.3×10^{-3}†
FCGR2B	rs4657041	1.0×10^{-3}	6.1×10^{-1}
CD40	rs4810485	1.5×10^{-3}	8.7×10^{-1}

Discussion

Proteome-by-phenome Mendelian Randomisation (P^2MR) is an efficient method of identifying potential drug targets through integrating pQTL with myriad phenotypes. P^2MR offers a data-driven approach to drug-discovery from population-level data. It quantifies the strength of evidence for causation, together with magnitude and direction of effect, for particular proteins in specific disease phenotypes. In addition, because MR using locally-acting pQTLs is more focussed than a genome-wide study, the burden of multiple testing is reduced dramatically, effectively reducing the sample-size required to declare a given effect significant.

P^2MR has some inherent limitations that need to be considered when interpreting results. First, a true positive MR association in our analysis implies that any intervention to replicate the effect of a given genotype would alter the relevant phenotype. Nevertheless, this association is informative neither of the time interval, during development for example, nor the anatomical location in which an intervention would need to be delivered. Second, pleiotropic effects cannot be excluded entirely without (unachievable) quantification of every mediator. Third, the concentration of a protein in plasma could be an imperfect proxy for the effect of a drug targeting that protein at the level of a whole organism. Finally, plasma concentration does not necessarily reflect activity. For example, a variant may cause expression of high levels of an inactive form of a protein. Or, for proteins with both membrane-bound and unbound forms, the MR direction of effect observed from quantifying soluble protein abundance may not reflect that of membrane-bound protein. For many membrane-bound proteins, a soluble (often antagonistic) form exists that is commonly produced through alternative splicing or proteolytic cleavage of the membrane-bound form. For example, tocilizumab, an IL6 receptor antagonist, is used as a treatment of rheumatoid arthritis (L. J. Scott, 2017). The variant we use to instrument IL6R level, rs61812598, is in complete LD with the missense variant rs2228145 in the British sub-population of 1,000 Genomes (Sudmant et al., 2015; The 1000 Genomes Project Consortium, 2015) whose effects on proteolytic cleavage of the membrane-bound form and alternative splicing have been

examined in detail (R. C. Ferreira et al., 2013). Carriers of the 358Ala allele at rs2228145 tend to have increased soluble IL6R but reduced membrane-bound IL6R in a number of immune cell types. Differences between the effects of soluble and membrane-bound forms of a protein may be wide-spread. For example, Dupilumab is a monoclonal antibody that targets IL4R, a key component of both IL4 and IL13 signalling. It is currently under investigation for the treatment of asthma and has shown promising results in both eosinophilic and non-eosinophilic asthma (Wenzel et al., 2016, 2013). Based on our results, we would have predicted that increased levels of IL4R result in a lower risk of asthma (Supplementary Table S3). This is in contrast to the direction-of-effect due to dupilumab administration. However, as with IL6R, IL4R has both a soluble and a membrane-bound form. Encouragingly, despite this, a relationship between dupilumab and asthma remains plausible – as evidenced by the 14 recently completed or ongoing clinical trials to assess the efficacy and safety of dupilumab in asthma (As of 26 Mar 2019, ClinicalTrials.gov Identifiers: NCT01312961, NCT01854047, NCT02134028, NCT02414854, NCT02528214, NCT02573233, NCT02948959, NCT03112577, NCT03387852, NCT03560466, NCT03620747, NCT03694158, NCT03782532, and NCT03884842).

P²MR provides an opportunity for studying the probable effects of specific proteins upon human diseases, such as schizophrenia, for which only imperfect model systems currently exist. Without a robust disease model, one must undertake studies in humans. However, there is little justification to undertake an adequately powered randomised control trial of a drug targeting a protein for which there is minimal evidence of a link between that protein and disease. P²MR does not suffer from such limitations.

P²MR highlights FABP2 as contributory to the pathogenesis of CAD and there are orthogonal lines of evidence to support this; notably: the non-synonymous mutation Ala54Thr (Yuan et al., 2015). In addition, given its interaction with PPAR-α and fenofibrate (Hughes et al., 2015) and strong expression in the gastrointestinal tract ("FABP2 available from

v18.proteinatlas.org," 2018; "The Human Protein Atlas," n.d.; Thul et al., 2017; Uhlén et al., 2015; Uhlen et al., 2017), FABP2 represents a potential drug-target of the future.

Finally, as well as its utility in identifying potential therapeutic targets for drug development, P²MR allows for an assessment of potential off-target effects of existing pharmacological targets. For example, we predict an effect of IL4R modulation on eosinophil count and percentage. This is an association already realised in one of the phase II clinical trials investigating dupilumab in asthma: a rise in eosinophil count was observed for some patients, even leading to the withdrawal of one patient from the study (Wenzel et al., 2016, 2013).

Conclusions

In summary, we have identified dozens of plausible causal links by conducting GWA of 249 proteins, followed by phenome-wide MR using replicated locally-acting pQTLs of 64 proteins: P²MR.

Using this approach, 54,144 protein-outcome links have been assessed and 509 significant (FDR <0.05) links identified: including anthropometric measures, haematological parameters, as well as diseases. Opportunities to discover larger sets of plausible causal links will increase as study sizes and pQTL numbers grow. Indeed, whole-proteome versus Biobank GWA Atlas studies will likely become feasible as pQTL measurement technologies mature.

Methods

Cohort description. From the islands of Orkney (Scotland) and Vis (Croatia) respectively, the ORCADES (McQuillan et al., 2008) and CROATIA-Vis (Campbell et al., 2007; Rudan et al., 2009) studies are of two isolated population cohorts that are both genotyped and richly phenotyped.

The Orkney Complex Disease Study (ORCADES) is a family-based, cross-sectional study that seeks to identify genetic factors influencing cardiovascular and other disease risk in the isolated archipelago of the Orkney Isles in northern Scotland (McQuillan et al., 2008). Genetic diversity in this population is decreased compared to Mainland Scotland, consistent with the high levels of endogamy historically. 2,078 participants aged 16-100 years were recruited between 2005 and 2011, most having three or four grandparents from Orkney, the remainder with two Orcadian grandparents. Fasting blood samples were collected and many health-related phenotypes and environmental exposures were measured in each individual. All participants gave written informed consent and the study was approved by Research Ethics Committees in Orkney and Aberdeen (North of Scotland REC).

The CROATIA-Vis study includes 1,008 Croatians, aged 18-93 years, who were recruited from the villages of Vis and Komiza on the Dalmatian island of Vis during spring of 2003 and 2004. All participants were volunteers and gave informed consent. They underwent a medical examination and interview, led by research teams from the Institute for Anthropological Research and the Andrija Stampar School of Public Health, (Zagreb, Croatia). All subjects visited the clinical research centre in the region, where they were examined in person and where fasting blood was drawn and stored for future analyses. Many biochemical and physiological measurements were performed, and questionnaires of medical history as well as lifestyle and environmental exposures were collected. The study received approval from the relevant ethics committees in Scotland and Croatia (REC reference: 11/AL/0222) and complied with the tenets of the Declaration of Helsinki.

Genotyping. Chromosomes and positions reported in this paper are from GRCh37 throughout. Genotyping of the ORCADES cohort was performed on the Illumina Human Hap 300v2, Illumina Omni Express, and Illumina Omni 1 arrays; that of the CROATIA-Vis cohort used the Illumina HumanHap300v1 array.

The genotyping array data were subject to the following quality control thresholds: genotype call-rate 0.98, per-individual call-rate 0.97, failed Hardy-Weinberg test at p-value $< 1 \times 10^{-6}$,

385 and minor allele frequency 0.01; genomic relationship matrix and principal components
386 were calculated using GenABEL (1.8-0) (Aulchenko, Ripke, Isaacs, & van Duijn, 2007) and
387 PLINK v1.90 (C. C. Chang et al., 2015; Purcell, 2017).
388
389 Assessment for ancestry outliers was performed by anchored PCA analysis when compared
390 to all non-European populations from the 1,000 Genomes project (Sudmant et al., 2015; The
391 1000 Genomes Project Consortium, 2015). Individuals with a mean-squared distance of
392 >10% in the first two principal components were removed. Genotypes were phased using
393 Shapeit v2.r873 and duoHMM (O'Connell et al., 2014) and imputed to the HRC.r1-1 reference
394 panel (The Haplotype Reference Consortium et al., 2016). 278,618 markers (Hap300) and
395 599,638 markers (Omni) were used for the imputation in ORCADES, and 272,930 markers
396 for CROATIA-Vis.
397
398 **Proteomics.** Plasma abundance of 249 proteins was measured in two European cohorts
399 using Olink Proseek Multiplex CVD2, CVD3, and INF panels. All proteomics measurements
400 were obtained from fasting EDTA plasma samples. Following quality control, there were 971
401 individuals in ORCADES, and 887 individuals in CROATIA-Vis, who had genotype and
402 proteomic data from Olink CVD2, 993 and 899 from Olink CVD3, and 982 and 894 from Olink
403 INF. The Olink Proseek Multiplex method uses a matched pair of antibodies for each protein,
404 linked to paired oligonucleotides. Binding of the antibodies to the protein brings the
405 oligonucleotides into close proximity and permits hybridization. Following binding and
406 extension, these oligonucleotides form the basis of a quantitative PCR reaction that allows
407 relative quantification of the initial protein concentration (Assarsson et al., 2014). Olink
408 panels include internal and external controls on each plate: two controls of the immunoassay
409 (two non-human proteins), one control of oligonucleotide extension (an antibody linked to
410 two matched oligonucleotides for immediate proximity, independent of antigen binding) and
411 one control of hybridized oligonucleotide detection (a pre-made synthetic double stranded
412 template), as well as an external, between-plate, control ("Olink," n.d.).
413

Prior to analysis, we excluded proteins with fewer than 200 samples with measurements above the limit of detection of the assay. Of the 268 unique proteins reported by Olink, 253 passed this threshold in ORCADES, and 252 in CROATIA-Vis, with an intersect of 251 proteins. Protein values were inverse-normal rank-transformed prior to subsequent analysis.

The subunits of IL27 are not distinguished in Olink's annotation (Q14213, *EBI3*; and Q8NEV9, *IL27*). However, it has only one significant locus, local to the *EBI3* gene (lead variant, rs60160662, is within 16kb). Therefore, *EBI3* (Q14213) was selected as representative for this protein when discussing pQTL location (local/distal) so as to avoid double counting.

Two proteins, CCL20 and BDNF, have been removed at the request of Olink.

Genome-wide association of protein levels. Genome-wide association of these proteins was performed using autosomes only. Analyses were performed in three-stages. (1) a linear regression model was used to account for participant age, sex, genotyping array (ORCADES only), proteomics plate, proteomics plate row, proteomics plate column, length of sample storage, season of venepuncture (ORCADES only), and the first 10 principal components of the genomic relationship matrix. Genotyping array and season of venepuncture are invariant in CROATIA-Vis and therefore were not included in the model. (2) Residuals from this model were corrected for relatedness, using GenABELs (Aulchenko et al., 2007) polygenic function and the genomic relationship matrix, to produce GRAMMAR+ residuals. Outlying GRAMMAR+ residuals (absolute z-score >4) were removed and the remainder rank-based inverse-normal transformed. (3) Genome-wide association testing was performed using REGSCAN v0.5 (Haller, Kals, Esko, Magi, & Fischer, 2013).

Reported pQTLs. Genome-wide association results were clumped by linkage disequilibrium using PLINK v1.90 (C. C. Chang et al., 2015; Purcell, 2017). Biallelic variants within ±5Mb and

443 r^2 >0.2 to the lead variant (smallest p-value at the locus) were clumped together, and the
444 lead variant is presented. r^2 was derived from all European populations in 1,000 Genomes
445 (Sudmant et al., 2015; The 1000 Genomes Project Consortium, 2015).

446

447 **Mendelian Randomisation.** In the context of P²MR, a DNA variant (a single nucleotide
448 polymorphism in this case) that influences plasma protein level is described as an
449 'instrumental variable', the protein as the 'exposure variable', and the outcome phenotype as
450 the 'outcome variable'.
451 A DNA variant was considered to be a potentially valid instrumental variable if it met the
452 following criteria:

453 (1) Minor allele frequency >1% in both ORCADES and CROATIA-Vis cohorts.
454 (2) An imputation info score (SNPTEST v2) of >0.95 in both ORCADES and CROATIA-Vis.
455 (3) Located within ±150kb of the gene coding for the protein (start and end coordinates
456 of the gene as defined by Ensembl GRCh37 (Zerbino et al., 2018)).

457

458 DNA variant selection was performed using the discovery (CROATIA-Vis) cohort. Replication
459 was defined based on a Bonferroni correction for the number of genome-wide significant
460 lead variants selected in the discovery cohort (CROATIA-Vis). In order to avoid a 'winner's
461 curse', replicated genome-wide association effect sizes and standard errors from the
462 replication cohort (ORCADES) were used for MR.

463

464 We perform MR as a ratio of expectations, using up to second-order partial derivatives of the
465 Taylor series expansion for effect size estimates, and up to first-order for standard errors
466 (Delta method) (Lynch & Walsh, 1998):

467

468

469

$$\beta_{YX} \approx \frac{\beta_{YZ}}{\beta_{XZ}} \left(1 + \frac{se_{XZ}^2}{\beta_{XZ}^2}\right) \quad (1)$$

$$se_{YX} \approx \sqrt{\frac{se_{YZ}^2}{\beta_{XZ}^2} + \frac{\beta_{YZ}^2 \cdot se_{XZ}^2}{\beta_{XZ}^4}} \quad (2)$$

$$p_{YX} = 2\Phi\left(-|\beta_{YX}|/se_{YX}\right) \quad (3)$$

where β_{ij} is the causal effect of *j* on *i*, se_{ij} is the standard error of the causal effect estimate of *j* on *i*; subscript *X* is the exposure, *Y* the outcome trait, and *Z* the instrumental variable. Φ is the cumulative density function of the standard normal distribution.

DNA variant to trait association: GeneAtlas. All outcome GWA (778 traits) from GeneAtlas (Canela-Xandri et al., 2017) were included. For each protein, the lead (lowest DNA variant-protein association p-value in the discovery cohort) biallelic (Phase 3, 1,000 Genomes (Sudmant et al., 2015; The 1000 Genomes Project Consortium, 2015)) variant meeting the criteria above and an imputation info score >0.95 in UK Biobank, was selected for each protein, and MR performed. An FDR of <0.05 was considered to be significant.

DNA variant to trait association: Phenoscanner. Phenoscanner ("PhenoScanner," 2018; Staley et al., 2016) was used to highlight existing GWA studies for inclusion. For each protein, the lead (lowest DNA variant-protein association p-value in the discovery cohort) biallelic (1,000 Genomes (Sudmant et al., 2015; The 1000 Genomes Project Consortium, 2015)) meeting the criteria above was selected. rs545634 was not found in the Phenoscanner database and was therefore replaced with the second most significant variant meeting the above criteria: chr1:15849003. Phenoscanner was run with the following options: Catalogue: 'Diseases & Traits', p-value cut-off: '1', Proxies: 'None', Build '37'. Results from 20 additional studies were obtained, corresponding to 68 outcomes. The results from those studies that returned a value for all input variants were kept and MR performed. An FDR of <0.05 was considered to be significant.

HEIDI. heterogeneity in dependent instruments (HEIDI) analysis (Zhu et al., 2016), is a method of testing whether the MR estimates obtained using variants in linkage disequilibrium with the lead variant are consistent with a single causal variant or multiple

497 causal variant at a given locus (Figure 1D). HEIDI analysis was performed using software
498 provided at https://cnsgenomics.com/software/smr/ [accessed 28/08/2018; v0.710]. We
499 created a bespoke BESD format file containing the pQTL data from ORCADES for assessment
500 as the exposure. Biallelic variants from the 1,000 Genomes (Sudmant et al., 2015; The 1000
501 Genomes Project Consortium, 2015) (European populations: CEU, FIN, GBR, IBS, and TSI)
502 were used as the linkage disequilibrium reference. We used the default 'cis-window' of
503 2000kb, and a maximum number of variants of 20 (as this is now the default value for the
504 software: based on unpublished power calculations by the authors of HEIDI and noted on
505 their website).
506
507 We performed HEIDI analysis of all exposure-outcome links that were found to be significant
508 (FDR <0.05) using outcomes from UK Biobank (n=271), as well as those links found to be MR
509 significant (FDR <0.05) with CAD from the meta-analysis of van der Harst (van der Harst &
510 Verweij, 2018), and for SHPS1 and schizophrenia (Schizophrenia Working Group of the
511 Psychiatric Genomics Consortium et al., 2014).
512
513 We applied the following filters for variants to be included in the analysis: minor allele
514 frequency MAF > 0.01 and, in the GeneAtlas and ORCADES data, an imputation info score of
515 >0.95.
516
517 **Schizophrenia GWA study replication.** In the initial analysis of Canela-Xandri et al. (2017),
518 schizophrenia was included as 'F20 Schizophrenia', and nested in 'F20-F29 Schizophrenia,
519 schizotypal and delusional disorders'. There were 920 cases in 'F20-F29 Schizophrenia,
520 schizotypal and delusional disorders' and 509 in 'F20 Schizophrenia'. Due to the near
521 doubling of the sample size, replication was attempted in the parent category: 'F20-F29
522 Schizophrenia, schizotypal and delusional disorders'. Using a Bonferroni correction, none of
523 these links replicated. However, due to the severe contraction of the number of cases present
524 in the sample – 35,476 cases and 46,839 controls to 920 cases and 407,535 controls – there
525 was a significant risk of false negative results. In order to address this, we re-analysed the UK

Biobank data including 'F20-F29 Schizophrenia, schizotypal and delusional disorders' and self-reported 'schizophrenia' as a single outcome in a more permissive set of individuals: individuals self-reporting their ethnicity as 'White' and clustering as a group based on the first two genomic principal components (Canela-Xandri, Rawlik, & Tenesa, 2018). This increased the number of cases and controls to 1,241 cases and 451,023 controls.

Acknowledgements

- A debt of gratitude is owed to all the participants in all cohorts used, without whom this work would not have been possible.
- This research has been conducted using the UK Biobank Resource under project 788.
- Funding:
  - ADB would like to acknowledge funding from the Wellcome PhD training fellowship for clinicians (204979/Z/16/Z), the Edinburgh Clinical Academic Track (ECAT) programme.
  - TB, YZ, CA, PN, JFW, VV, CHay, CPP and CHal are supported by MRC University Unit Programme Grants to the Human Genetics Unit (MC_PC_U127592696, MC_UU_12008/1, MC_UU_00007/10 and MC_UU_00007/15)
  - AT, OC-X and KR acknowledge funding from the MRC (MR/R025851/1, MR/N003179/1).
  - CHal, JKB, AT, and KR acknowledge funding from BBSRC Institute Strategic Programme grants to the Roslin Institute (BBS/E/D/30002275, BBS/E/D/30002276, BBS/E/D/10002071, BBS/E/D/20002172, BBS/E/D/20002174).
  - PKJ would like to acknowledge funding from the Axa research fund.

- JKB acknowledges funding support from a Wellcome-Beit Prize Intermediate Clinical Fellowship (103258/Z/13/Z,A), and the UK Intensive Care Foundation.

- The Orkney Complex Disease Study (ORCADES) was supported by the Chief Scientist Office of the Scottish Government (CZB/4/276, CZB/4/710), a Royal Society URF to JFW, the MRC Human Genetics Unit quinquennial programme "QTL in Health and Disease", Arthritis Research UK and the European Union framework program 6 EUROSPAN project (contract no. LSHG-CT-2006-018947). DNA extractions were performed at the Wellcome Trust Clinical Research Facility in Edinburgh. We would like to acknowledge the invaluable contributions of the research nurses in Orkney, the administrative team in Edinburgh and the people of Orkney.
- The CROATIA-Vis study was funded by grants from the Medical Research Council (UK) and Republic of Croatia Ministry of Science, Education and Sports research grants. (108-1080315-0302). We would like to acknowledge the staff of several institutions in Croatia that supported the field work, including but not limited to The University of Split and Zagreb Medical Schools, the Institute for Anthropological Research in Zagreb and Croatian Institute for Public Health. Genotyping was performed in the Genetics Core of the Clinical Research Facility, University of Edinburgh.

References

Arrowsmith, J. (2011). Trial watch: Phase II failures: 2008-2010. *Nature Reviews Drug Discovery*, *10*, 328–329. https://doi.org/10.1038/nrd3439

Assarsson, E., Lundberg, M., Holmquist, G., Björkesten, J., Bucht Thorsen, S., Ekman, D., … Fredriksson, S. (2014). Homogenous 96-Plex PEA Immunoassay Exhibiting High Sensitivity, Specificity, and Excellent Scalability. *PLOS ONE*, *9*(4), e95192. https://doi.org/10.1371/journal.pone.0095192

Astle, W. J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A. L., … Soranzo, N. (2016). The Allelic Landscape of Human Blood Cell Trait Variation and

582 Links to Common Complex Disease. *Cell*, *167*(5), 1415-1429.e19.
583 https://doi.org/10.1016/j.cell.2016.10.042
584 Aulchenko, Y. S., Ripke, S., Isaacs, A., & van Duijn, C. M. (2007). GenABEL: an R
585 library for genome-wide association analysis. *Bioinformatics (Oxford,*
586 *England)*, *23*(10), 1294–1296.
587 https://doi.org/10.1093/bioinformatics/btm108
588 Baillie, J. K. (2014). Translational genomics. Targeting the host immune response to
589 fight infection. *Science (New York, N.Y.)*, *344*(6186), 807–808.
590 https://doi.org/10.1126/science.1255074
591 Baillie, J. K., Bretherick, A., Haley, C. S., Clohisey, S., Gray, A., Neyton, L. P. A., … Hume,
592 D. A. (2018). Shared activity patterns arising at genetic susceptibility loci
593 reveal underlying genomic and cellular architecture of human disease. *PLoS*
594 *Computational Biology*, *14*(3), e1005934.
595 https://doi.org/10.1371/journal.pcbi.1005934
596 Beaumont, R. N., Warrington, N. M., Cavadino, A., Tyrrell, J., Nodzenski, M.,
597 Horikoshi, M., … Freathy, R. M. (2018). Genome-wide association study of
598 offspring birth weight in 86 577 women identifies five novel loci and
599 highlights maternal genetic effects that are independent of fetal genetics.
600 *Human Molecular Genetics*, *27*(4), 742–756.
601 https://doi.org/10.1093/hmg/ddx429
602 Berg, S. M. van den, Moor, M. H. M. de, Verweij, K. J. H., Krueger, R. F., Luciano, M.,
603 Vasquez, A. A., … Boomsma, D. I. (2016). Meta-analysis of Genome-Wide
604 Association Studies for Extraversion: Findings from the Genetics of
605 Personality Consortium. *Behavior Genetics*, *46*(2), 170–182.
606 https://doi.org/10.1007/s10519-015-9735-5
607 Bronson, P. G., Chang, D., Bhangale, T., Seldin, M. F., Ortmann, W., Ferreira, R. C., …
608 Behrens, T. W. (2016). Common variants at PVT1, ATG13-AMBRA1, AHI1
609 and CLEC16A are associated with selective IgA deficiency. *Nature Genetics*,
610 *48*(11), 1425–1429. https://doi.org/10.1038/ng.3675

611 Brown, G. C., & Neher, J. J. (2014). Microglial phagocytosis of live neurons. *Nature*
612 *Reviews Neuroscience*, *15*(4), 209–216. https://doi.org/10.1038/nrn3710
613 Campbell, H., Carothers, A. D., Rudan, I., Hayward, C., Biloglav, Z., Barac, L., … Wright,
614 A. F. (2007). Effects of genome-wide heterozygosity on a range of
615 biomedically relevant human quantitative traits. *Human Molecular Genetics*,
616 *16*(2), 233–241. https://doi.org/10.1093/hmg/ddl473
617 Canela-Xandri, O., Rawlik, K., & Tenesa, A. (2017). An atlas of genetic associations in
618 UK Biobank. *BioRxiv*, 176834. https://doi.org/10.1101/176834
619 Canela-Xandri, O., Rawlik, K., & Tenesa, A. (2018). An atlas of genetic associations in
620 UK Biobank. *Nature Genetics*, *50*(11), 1593.
621 https://doi.org/10.1038/s41588-018-0248-z
622 Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015).
623 Second-generation PLINK: rising to the challenge of larger and richer
624 datasets. *GigaScience*, *4*(1), 1–16. https://doi.org/10.1186/s13742-015-
625 0047-8
626 Chang, H. P., Lindberg, F. P., Wang, H. L., Huang, A. M., & Lee, E. H. Y. (1999).
627 Impaired Memory Retention and Decreased Long-Term Potentiation in
628 Integrin-Associated Protein-Deficient Mice. *Learning & Memory*, *6*(5), 448–
629 457.
630 Civelek, M., & Lusis, A. J. (2014). Systems genetics approaches to understand
631 complex traits. *Nature Reviews Genetics*, *15*(1), 34–48.
632 https://doi.org/10.1038/nrg3575
633 de Moor, M. H. M., van den Berg, S. M., Verweij, K. J. H., Krueger, R. F., Luciano, M.,
634 Arias Vasquez, A., … Boomsma, D. I. (2015). Meta-analysis of Genome-wide
635 Association Studies for Neuroticism, and the Polygenic Association With
636 Major Depressive Disorder. *JAMA Psychiatry*, *72*(7), 642.
637 https://doi.org/10.1001/jamapsychiatry.2015.0554

638 FABP2 available from v18.proteinatlas.org. (2018, November 15). Retrieved April 1,
639 2019, from The Human Protein Atlas website:
640 https://v18.proteinatlas.org/ENSG00000145384-FABP2/tissue
641 Ferreira, M. A., Vonk, J. M., Baurecht, H., Marenholz, I., Tian, C., Hoffman, J. D., …
642 Paternoster, L. (2017). Shared genetic origin of asthma, hay fever and
643 eczema elucidates allergic disease biology. *Nature Genetics*, *49*(12), 1752–
644 1757. https://doi.org/10.1038/ng.3985
645 Ferreira, R. C., Freitag, D. F., Cutler, A. J., Howson, J. M. M., Rainbow, D. B., Smyth, D. J.,
646 … Todd, J. A. (2013). Functional IL6R 358Ala Allele Impairs Classical IL-6
647 Receptor Signaling and Influences Risk of Diverse Inflammatory Diseases.
648 *PLOS Genetics*, *9*(4), e1003444.
649 https://doi.org/10.1371/journal.pgen.1003444
650 Finan, C., Gaulton, A., Kruger, F. A., Lumbers, R. T., Shah, T., Engmann, J., … Casas, J. P.
651 (2017). The druggable genome and support for target identification and
652 validation in drug development. *Science Translational Medicine*, *9*(383),
653 eaag1166. https://doi.org/10.1126/scitranslmed.aag1166
654 Folkersen, L., Fauman, E., Sabater-Lleal, M., Strawbridge, R. J., Frånberg, M.,
655 Sennblad, B., … Mälarstig, A. (2017). Mapping of 79 loci for 83 plasma
656 protein biomarkers in cardiovascular disease. *PLOS Genetics*, *13*(4),
657 e1006706. https://doi.org/10.1371/journal.pgen.1006706
658 Haller, T., Kals, M., Esko, T., Magi, R., & Fischer, K. (2013). RegScan: a GWAS tool for
659 quick estimation of allele effects on continuous traits and their
660 combinations. *Briefings in Bioinformatics*, *16*(1), 39–44.
661 https://doi.org/10.1093/bib/bbt066
662 Hammerschlag, A. R., Stringer, S., de Leeuw, C. A., Sniekers, S., Taskesen, E.,
663 Watanabe, K., … Posthuma, D. (2017). Genome-wide association analysis of
664 insomnia complaints identifies risk genes and genetic overlap with
665 psychiatric and metabolic traits. *Nature Genetics*, *49*(11), 1584–1592.
666 https://doi.org/10.1038/ng.3888

667 Hou, L., Bergen, S. E., Akula, N., Song, J., Hultman, C. M., Landén, M., … McMahon, F. J.
668 (2016). Genome-wide association study of 40,000 individuals identifies two
669 novel loci associated with bipolar disorder. *Human Molecular Genetics*,
670 *25*(15), 3383–3394. https://doi.org/10.1093/hmg/ddw181
671 Huang, A.-M., Wang, H. L., Tang, Y. P., & Lee, E. H. Y. (1998). Expression of Integrin-
672 Associated Protein Gene Associated with Memory Formation in Rats.
673 *Journal of Neuroscience*, *18*(11), 4305–4313.
674 Hughes, M. L. R., Liu, B., Halls, M. L., Wagstaff, K. M., Patil, R., Velkov, T., … Porter, C. J.
675 H. (2015). Fatty Acid-binding Proteins 1 and 2 Differentially Modulate the
676 Activation of Peroxisome Proliferator-activated Receptor α in a Ligand-
677 selective Manner. *The Journal of Biological Chemistry*, *290*(22), 13895–
678 13906. https://doi.org/10.1074/jbc.M114.605998
679 IL6R Genetics Consortium Emerging Risk Factors Collaboration. (2012).
680 Interleukin-6 receptor pathways in coronary heart disease: a collaborative
681 meta-analysis of 82 studies. *The Lancet*, *379*(9822), 1205–1213.
682 https://doi.org/10.1016/S0140-6736(11)61931-4
683 Interleukin-6 Receptor Mendelian Randomisation Analysis (IL6R MR) Consortium,
684 Swerdlow, D. I., Holmes, M. V., Kuchenbaecker, K. B., Engmann, J. E. L., Shah,
685 T., … Casas, J. P. (2012). The interleukin-6 receptor as a target for
686 prevention of coronary heart disease: a mendelian randomisation analysis.
687 *Lancet (London, England)*, *379*(9822), 1214–1224.
688 https://doi.org/10.1016/S0140-6736(12)60110-X
689 Klarin, D., Zhu, Q. M., Emdin, C. A., Chaffin, M., Horner, S., McMillan, B. J., …
690 Kathiresan, S. (2017). Genetic Analysis in UK Biobank Links Insulin
691 Resistance and Transendothelial Migration Pathways to Coronary Artery
692 Disease. *Nature Genetics*, *49*(9), 1392–1397.
693 https://doi.org/10.1038/ng.3914
694 Koshimizu, H., Takao, K., Matozaki, T., Ohnishi, H., & Miyakawa, T. (2014).
695 Comprehensive Behavioral Analysis of Cluster of Differentiation 47

Knockout Mice. *PLOS ONE*, *9*(2), e89584. https://doi.org/10.1371/journal.pone.0089584

Liu, J. Z., van Sommeren, S., Huang, H., Ng, S. C., Alberts, R., Takahashi, A., … Weersma, R. K. (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature Genetics*, *47*(9), 979–986. https://doi.org/10.1038/ng.3359

Lynch, M., & Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits* (1998 edition). Sunderland, Mass: Sinauer.

Ma, L., Kulesskaya, N., Võikar, V., & Tian, L. (2015). Differential expression of brain immune genes and schizophrenia-related behavior in C57BL/6N and DBA/2J female mice. *Psychiatry Research*, *226*(1), 211–216. https://doi.org/10.1016/j.psychres.2015.01.001

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., … Parkinson, H. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, *45*(Database issue), D896–D901. https://doi.org/10.1093/nar/gkw1133

Martins-de-Souza, D., Gattaz, W. F., Schmitt, A., Rewerts, C., Maccarrone, G., Dias-Neto, E., & Turck, C. W. (2009). Prefrontal cortex shotgun proteome analysis reveals altered calcium homeostasis and immune system imbalance in schizophrenia. *European Archives of Psychiatry and Clinical Neuroscience*, *259*(3), 151–163. https://doi.org/10.1007/s00406-008-0847-2

McQuillan, R., Leutenegger, A.-L., Abdel-Rahman, R., Franklin, C. S., Pericic, M., Barac-Lauc, L., … Wilson, J. F. (2008). Runs of homozygosity in European populations. *American Journal of Human Genetics*, *83*(3), 359–372. https://doi.org/10.1016/j.ajhg.2008.08.007

723 Mirauta, B., Seaton, D. D., Bensaddek, D., Brenes, A., Bonder, M. J., Kilpinen, H., …
724 Lamond, A. (2018). Population-scale proteome variation in human induced
725 pluripotent stem cells. *BioRxiv*, 439216. https://doi.org/10.1101/439216
726 Mouse Genome Informatics (v6.13). (2019, March 26). Retrieved April 1, 2019, from
727 http://www.informatics.jax.org/
728 Munos, B. (2009). Lessons from 60 years of pharmaceutical innovation. *Nature*
729 *Reviews. Drug Discovery*, *8*(12), 959–968.
730 https://doi.org/10.1038/nrd2961
731 Navarini, A. A., French, L. E., & Hofbauer, G. F. L. (2011). Interrupting IL-6–receptor
732 signaling improves atopic dermatitis but associates with bacterial
733 superinfection. *Journal of Allergy and Clinical Immunology*, *128*(5), 1128–
734 1130. https://doi.org/10.1016/j.jaci.2011.09.009
735 Nelson, C. P., Goel, A., Butterworth, A. S., Kanoni, S., Webb, T. R., Marouli, E., …
736 Deloukas, P. (2017). Association analyses based on false discovery rate
737 implicate new loci for coronary artery disease. *Nature Genetics*, *49*(9),
738 1385–1391. https://doi.org/10.1038/ng.3913
739 Nelson, M. R., Tipney, H., Painter, J. L., Shen, J., Nicoletti, P., Shen, Y., … Sanseau, P.
740 (2015). The support of human genetic evidence for approved drug
741 indications. *Nature Genetics*, *47*(8), 856–860.
742 https://doi.org/10.1038/ng.3314
743 O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., … Marchini, J.
744 (2014). A General Approach for Haplotype Phasing across the Full
745 Spectrum of Relatedness. *PLOS Genetics*, *10*(4), e1004234.
746 https://doi.org/10.1371/journal.pgen.1004234
747 Ohnishi, H., Kaneko, Y., Okazawa, H., Miyashita, M., Sato, R., Hayashi, A., … Matozaki,
748 T. (2005). Differential Localization of Src Homology 2 Domain-Containing
749 Protein Tyrosine Phosphatase Substrate-1 and CD47 and Its Molecular
750 Mechanisms in Cultured Hippocampal Neurons. *Journal of Neuroscience*,
751 *25*(10), 2702–2711. https://doi.org/10.1523/JNEUROSCI.5173-04.2005

752 Ohnishi, H., Murata, T., Kusakari, S., Hayashi, Y., Takao, K., Maruyama, T., … Matozaki,
753 T. (2010). Stress-Evoked Tyrosine Phosphorylation of Signal Regulatory
754 Protein α Regulates Behavioral Immobility in the Forced Swim Test. *Journal*
755 *of Neuroscience*, *30*(31), 10472–10483.
756 https://doi.org/10.1523/JNEUROSCI.0257-10.2010
757 Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., … Plenge, R. M. (2014).
758 Genetics of rheumatoid arthritis contributes to biology and drug discovery.
759 *Nature*, *506*(7488), 376–381. https://doi.org/10.1038/nature12873
760 Okbay, A., Beauchamp, J. P., Fontana, M. A., Lee, J. J., Pers, T. H., Rietveld, C. A., …
761 Benjamin, D. J. (2016). Genome-wide association study identifies 74 loci
762 associated with educational attainment. *Nature*, *533*(7604), 539–542.
763 https://doi.org/10.1038/nature17671
764 Olink. (n.d.). Retrieved June 19, 2016, from http://www.olink.com/
765 Ozaki, K., Ohnishi, Y., Iida, A., Sekine, A., Yamada, R., Tsunoda, T., … Tanaka, T.
766 (2002). Functional SNPs in the lymphotoxin-alpha gene that are associated
767 with susceptibility to myocardial infarction. *Nature Genetics*, *32*(4), 650–
768 654. https://doi.org/10.1038/ng1047
769 Phelan, C. M., Kuchenbaecker, K. B., Tyrer, J. P., Kar, S. P., Lawrenson, K., Winham, S.
770 J., … Pharoah, P. D. P. (2017). Identification of 12 new susceptibility loci for
771 different histotypes of epithelial ovarian cancer. *Nature Genetics*, *49*(5),
772 680–691. https://doi.org/10.1038/ng.3826
773 PhenoScanner. (2018). Retrieved September 25, 2018, from
774 http://www.phenoscanner.medschl.cam.ac.uk/
775 Pruim, R. J., Welch, R. P., Sanna, S., Teslovich, T. M., Chines, P. S., Gliedt, T. P., …
776 Willer, C. J. (2010). LocusZoom: regional visualization of genome-wide
777 association scan results. *Bioinformatics (Oxford, England)*, *26*(18), 2336–
778 2337. https://doi.org/10.1093/bioinformatics/btq419
779 Purcell, S. (2017). PLINK: v1.90b3.46 (Version v1.90).

780 Rudan, I., Marusić, A., Janković, S., Rotim, K., Boban, M., Lauc, G., … Polasek, O.
781 (2009). "10001 Dalmatians:" Croatia launches its national biobank.
782 *Croatian Medical Journal*, *50*(1), 4–6.
783 Schizophrenia Working Group of the Psychiatric Genomics Consortium, Ripke, S.,
784 Neale, B. M., Corvin, A., Walters, J. T. R., Farh, K.-H., … O'Donovan, M. C.
785 (2014). Biological insights from 108 schizophrenia-associated genetic loci.
786 *Nature*, *511*(7510), 421–427. https://doi.org/10.1038/nature13595
787 Scott, L. J. (2017). Tocilizumab: A Review in Rheumatoid Arthritis. *Drugs*, *77*(17),
788 1865–1879. https://doi.org/10.1007/s40265-017-0829-7
789 Scott, R. A., Scott, L. J., Mägi, R., Marullo, L., Gaulton, K. J., Kaakinen, M., … DIAbetes
790 Genetics Replication And Meta-analysis (DIAGRAM) Consortium. (2017).
791 An Expanded Genome-Wide Association Study of Type 2 Diabetes in
792 Europeans. *Diabetes*, *66*(11), 2888–2902. https://doi.org/10.2337/db16-
793 1253
794 SIRPA available from v18.proteinatlas.org. (2018, November 15). Retrieved April 1,
795 2019, from The Human Protein Atlas website:
796 https://v18.proteinatlas.org/ENSG00000198053-SIRPA/tissue
797 Smith, G. D., & Ebrahim, S. (2003). "Mendelian randomization": can genetic
798 epidemiology contribute to understanding environmental determinants of
799 disease? *International Journal of Epidemiology*, *32*(1), 1–22.
800 Sniekers, S., Stringer, S., Watanabe, K., Jansen, P. R., Coleman, J. R. I., Krapohl, E., …
801 Posthuma, D. (2017). Genome-wide association meta-analysis of 78,308
802 individuals identifies new loci and genes influencing human intelligence.
803 *Nature Genetics*, *49*(7), 1107–1112. https://doi.org/10.1038/ng.3869
804 Staley, J. R., Blackshaw, J., Kamat, M. A., Ellis, S., Surendran, P., Sun, B. B., …
805 Butterworth, A. S. (2016). PhenoScanner: a database of human genotype–
806 phenotype associations. *Bioinformatics*, *32*(20), 3207–3209.
807 https://doi.org/10.1093/bioinformatics/btw373

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., … Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, *526*(7571), 75–81. https://doi.org/10.1038/nature15394

Suhre, K., Arnold, M., Bhagwat, A. M., Cotton, R. J., Engelke, R., Raffler, J., … Graumann, J. (2017). Connecting genetic risk to disease end points through the human blood plasma proteome. *Nature Communications*, *8*, 14357. https://doi.org/10.1038/ncomms14357

Sun, B. B., Maranville, J. C., Peters, J. E., Stacey, D., Staley, J. R., Blackshaw, J., … Butterworth, A. S. (2018). Genomic atlas of the human plasma proteome. *Nature*, *558*(7708), 73. https://doi.org/10.1038/s41586-018-0175-2

The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

The CARDIoGRAMplusC4D Consortium, Nikpay, M., Goel, A., Won, H.-H., Hall, L. M., Willenborg, C., … Farrall, M. (2015). A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics*, *47*(10), 1121–1130. https://doi.org/10.1038/ng.3396

The EArly Genetics and Lifecourse Epidemiology (EAGLE) Eczema Consortium, Paternoster, L., Standl, M., Waage, J., Baurecht, H., Hotze, M., … Weidinger, S. (2015). Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. *Nature Genetics*, *47*(12), 1449–1456. https://doi.org/10.1038/ng.3424

The Haplotype Reference Consortium, McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., … Marchini, J. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, *48*(10), 1279–1283. https://doi.org/10.1038/ng.3643

The Human Protein Atlas. (n.d.). Retrieved December 13, 2017, from https://www.proteinatlas.org/

837 Thul, P. J., Åkesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Blal, H. A., …
838 Lundberg, E. (2017). A subcellular map of the human proteome. *Science*,
839 *356*(6340), eaal3321. https://doi.org/10.1126/science.aal3321
840 Toth, A. B., Terauchi, A., Zhang, L. Y., Johnson-Venkatesh, E. M., Larsen, D. J., Sutton,
841 M. A., & Umemori, H. (2013). Synapse maturation by activity-dependent
842 ectodomain shedding of SIRPα. *Nature Neuroscience*, *16*(10), 1417.
843 https://doi.org/10.1038/nn.3516
844 Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., …
845 Pontén, F. (2015). Tissue-based map of the human proteome. *Science*,
846 *347*(6220), 1260419. https://doi.org/10.1126/science.1260419
847 Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., … Ponten, F.
848 (2017). A pathology atlas of the human cancer transcriptome. *Science*,
849 *357*(6352), eaan2507. https://doi.org/10.1126/science.aan2507
850 Ullah, M. A., Sukkar, M., Ferreira, M., & Phipps, S. (2014). 53: IL-6R blockade: A new
851 personalised treatment for asthma? *Cytokine*, *70*(1), 40.
852 https://doi.org/10.1016/j.cyto.2014.07.060
853 van der Harst, P., & Verweij, N. (2018). Identification of 64 Novel Genetic Loci
854 Provides an Expanded View on the Genetic Architecture of Coronary Artery
855 Disease. *Circulation Research*, *122*(3), 433–443.
856 https://doi.org/10.1161/CIRCRESAHA.117.312086
857 van Rheenen, W., Shatunov, A., Dekker, A. M., McLaughlin, R. L., Diekstra, F. P., Pulit,
858 S. L., … Veldink, J. H. (2016). Genome-wide association analyses identify
859 new risk variants and the genetic architecture of amyotrophic lateral
860 sclerosis. *Nature Genetics*, *48*(9), 1043–1048.
861 https://doi.org/10.1038/ng.3622
862 Wenzel, S., Castro, M., Corren, J., Maspero, J., Wang, L., Zhang, B., … Teper, A. (2016).
863 Dupilumab efficacy and safety in adults with uncontrolled persistent
864 asthma despite use of medium-to-high-dose inhaled corticosteroids plus a
865 long-acting β2 agonist: a randomised double-blind placebo-controlled

pivotal phase 2b dose-ranging trial. *The Lancet*, *388*(10039), 31–44. https://doi.org/10.1016/S0140-6736(16)30307-5

Wenzel, S., Ford, L., Pearlman, D., Spector, S., Sher, L., Skobieranda, F., … Pirozzi, G. (2013). Dupilumab in Persistent Asthma with Elevated Eosinophil Levels. *New England Journal of Medicine*, *368*(26), 2455–2466. https://doi.org/10.1056/NEJMoa1304048

Yao, C., Chen, G., Song, C., Keefe, J., Mendelson, M., Huan, T., … Levy, D. (2018). Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nature Communications*, *9*(1), 3268. https://doi.org/10.1038/s41467-018-05512-x

Yuan, D., Yu, C., & Zeng, C. (2015). Association of I-FABP gene polymorphism and the risk of coronary heart disease. *International Journal of Clinical and Experimental Medicine*, *8*(9), 16389–16393.

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., … Flicek, P. (2018). Ensembl 2018. *Nucleic Acids Research*, *46*(D1), D754–D761. https://doi.org/10.1093/nar/gkx1098

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., … Yang, J. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics*, *48*(5), 481–487. https://doi.org/10.1038/ng.3538

887 **Supplementary Materials**

888 • Table S1. Additional studies identified using Phenoscanner: additional_studies.tsv.

889 • Table S2. Complete list of pQTLs (linkage disequilibrium clumped): indep_pqtl.tsv.

890 • Table S3. Mendelian Randomisation results from UK Biobank: df_ukbb_heidi.tsv.

891 • Table S4. Mendelian Randomisation results from studies identified using
892 Phenoscanner: df_phenoscanner.tsv.

893 • Figure S1. Significant (FDR <0.05) P2MR protein-outcome causal inferences:
894 haematology count subset.

895 • Figure S2. Significant (FDR <0.05) P2MR protein-outcome causal inferences:
896 haematology percentage subset.

897 • Figure S3. Significant (FDR <0.05) P2MR protein-outcome causal inferences:
898 haematology (non-count, non-percentage) subset.

899 • Figure S4. Significant (FDR <0.05) P2MR protein-outcome causal inferences:
900 anthropometric measurements subset.

901

902 **Figure 1**. Proteome-by-phenome Mendelian Randomisation (P^2MR).

903 A) Genome-wide associations of the plasma concentrations of 249 proteins from two
904 independent European cohorts (discovery and replication) were calculated. The plot
905 shows pQTL position against chromosomal location of the gene that encodes the protein
906 under study for all replicated pQTLs. The area of a filled circle is proportional to its -
907 log10(p-value) in the replication cohort. Blue circles indicate pQTLs ±150kb of the gene
908 ('local-pQTLs'); red circles indicate pQTLs more than 150kb from the gene. B, C) Local-
909 pQTLs of 64 proteins were taken forward for P^2MR analysis. These were assessed against
910 778 outcome phenotypes from GeneAtlas (Canela-Xandri et al., 2017) (panel B; UK
911 Biobank) and 68 phenotypes identified using Phenoscanner (Staley et al., 2016) (panel C).
912 In each set of results an FDR of <0.05 was considered significant. D) Heterogeneity in
913 dependent instruments (HEIDI (Zhu et al., 2016)) testing was undertaken for MR
914 significant results from UK Biobank (n = 271). This test seeks to distinguish a single causal
915 variant at a locus effecting both exposure and outcome directly (as in i) or in a causal chain
916 (as in ii), from two causal variants in linkage disequilibrium (as in iii), one effecting the
917 exposure and the other effecting the outcome.
918

Figure 1.

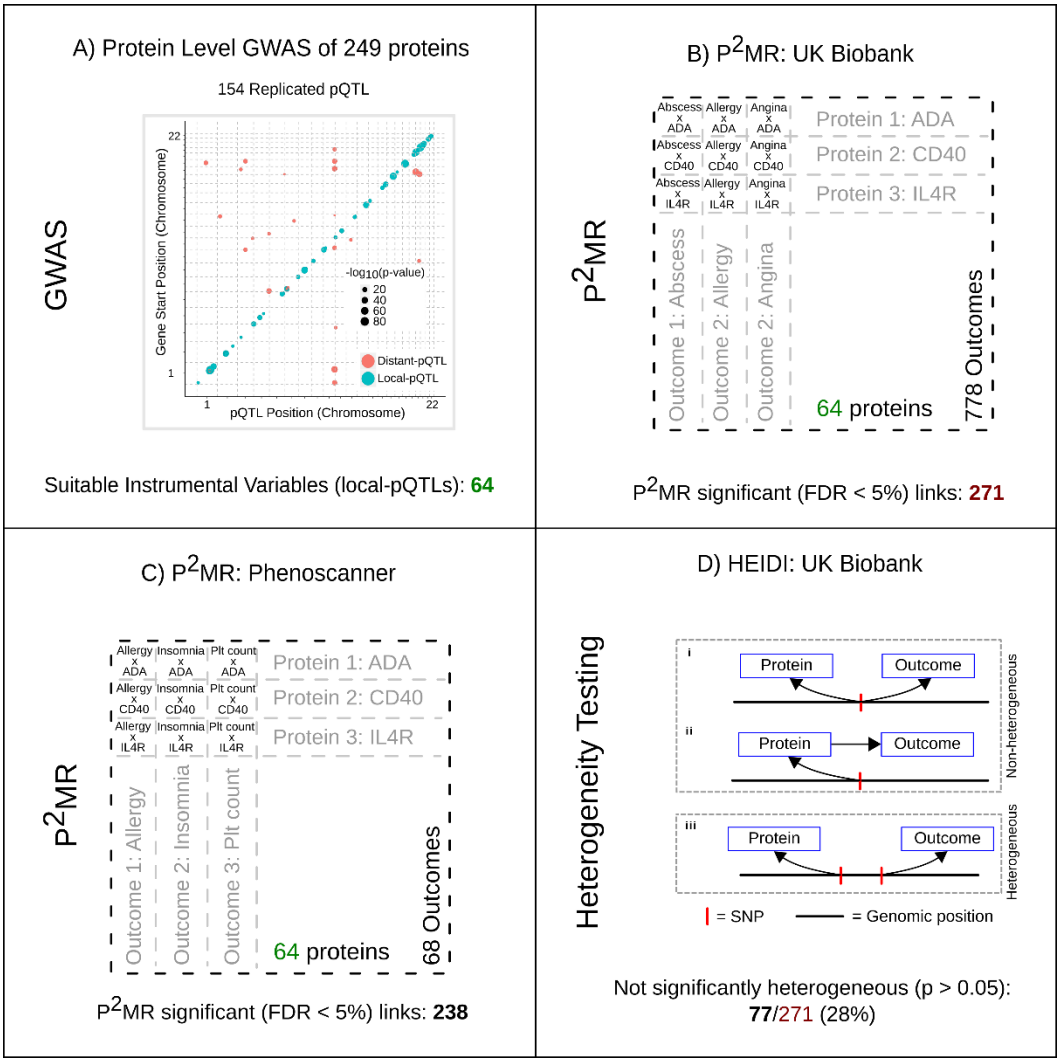
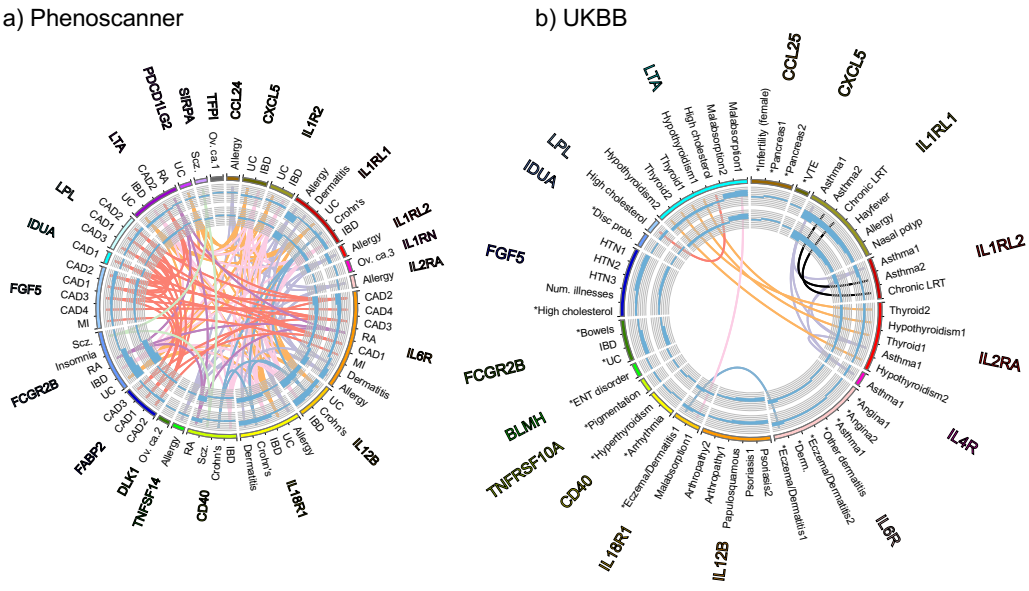


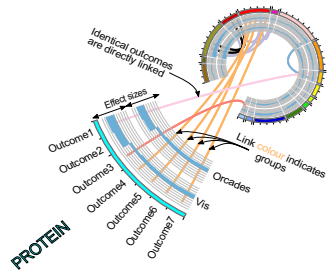
Figure 2. Significant (FDR <0.05) P²MR protein-outcome causal inferences: disease subset.

a) Phenoscanner (Staley et al., 2016): P²MR significant protein-disease outcome causal inferences for 20 Phenoscanner studies. b) GeneAtlas (Canela-Xandri et al., 2017): MR significant protein-disease outcome causal inferences for UK Biobank data. Asterisks indicate P²MR estimates that are not significantly heterogeneous (HEIDI, Main Text (Zhu et al., 2016)). Graphical key: Reading from the outside in: protein (exposure; HGNC symbol); disease outcome; key colour; bar chart of the signed (beta/standard error)^2 value of the MR estimate (using pQTL data from the discovery cohort; Methods); and bar chart of the signed (beta/standard error)^2 value of the MR estimate (using pQTL data from the replication cohort; Methods). Central chords join identical outcomes. Identically coloured chords indicate similar outcome groups, e.g. thyroid disease.

934 **Figure 2**



Key



Phenoscanner

Abbreviated Outcome	Outcome Description
Allergy	Allergic disease
CAD1	CAD (van der Harst et al. no UKBB)
CAD2	CAD (van der Harst et al. with UKBB)
CAD3	CAD (Nelson et al.)
CAD4	CAD (CARDIoGRAMplusC4D)
Crohn's	Crohns disease
Dermatitis	Atopic dermatitis
IBD	Inflammatory bowel disease
Insomnia	Insomnia
MI	MI (CARDIoGRAMplusC4D)
Ov. ca.1	Clear cell ovarian cancer
Ov. ca.2	Invasive ovarian cancer
Ov. ca.3	Mucinous ovarian cancer
RA	Rheumatoid arthritis
Scz.	Schizophrenia
UC	Ulcerative colitis

UKBB

Abbreviated Outcome	Outcome Description
Allergy	Allergy/hypersensitivity/anaphylaxis (self-reported)
Angina1	Angina (self-reported)
Angina2	I20 Angina pectoris
Arrhythmia	Heart arrhythmia (self-reported)
Arthropathy1	Psoriatic arthropathy (self-reported)
Arthropathy2	M07 Psoriatic and enteropathic arthropathies
Asthma1	Asthma (self-reported)
Asthma2	J45 Asthma
Bowels	Bowel problem (self-reported)
Chronic LRT	J40-J47 Chronic lower respiratory diseases
Derm.	Dermatology (self-reported)
Disc prob.	Disc problem (self-reported)
Eczema/Dermatitis1	Eczema/dermatitis (self-reported)
Eczema/Dermatitis2	L20-L30 Dermatitis and eczema
ENT disorder	Throat or larynx disorder (self-reported)
Hayfever	Hayfever/allergic rhinitis (self-reported)
High cholesterol	High cholesterol (self-reported)
HTN1	Hypertension (self-reported)
HTN2	I10-I15 Hypertensive diseases
HTN3	I10 Essential (primary) hypertension

UKBB

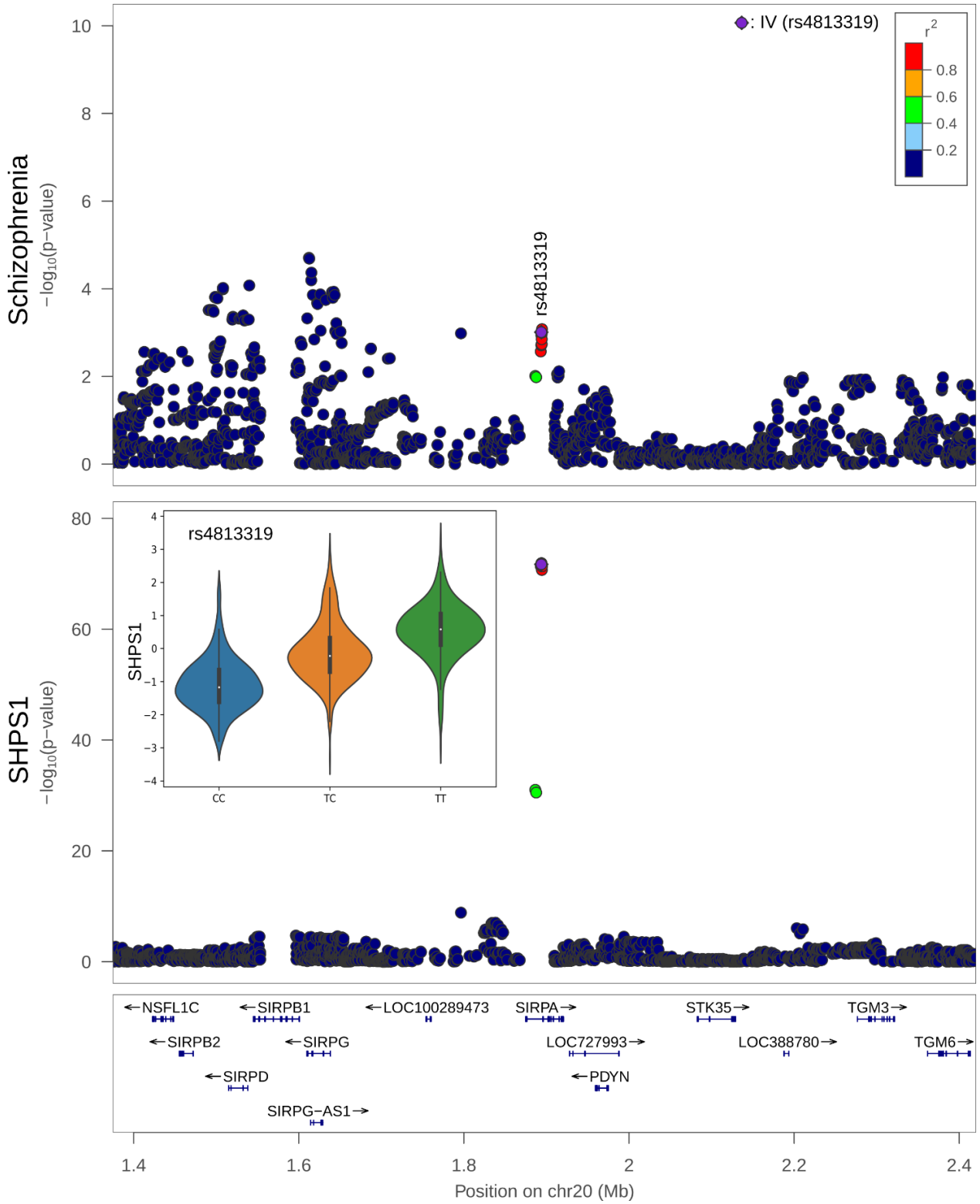
Abbreviated Outcome	Outcome Description
Hyperthyroidism	Hyperthyroidism/thyrotoxicosis (self-reported)
Hypothyroidism1	Hypothyroidism/myxoedema (self-reported)
Hypothyroidism2	E03 Other hypothyroidism
IBD	Inflammatory bowel disease (self-reported)
Infertility (female)	N97 Female infertility
Malabsorption1	Malabsorption/coeliac disease (self-reported)
Malabsorption2	K90 Intestinal malabsorption
Nasal polyp	J33 Nasal polyp
Num. illnesses	Number of self-reported non-cancer illnesses
Other dermatitis	L30 Other dermatitis
Pancreas1	E15-E16 Other disorders of glucose regulation and pancreatic internal secretion
Pancreas2	E16 Other disorders of pancreatic internal secretion
Papulosquamous	L40-L45 Papulosquamous disorders
Pigmentation	L81 Other disorders of pigmentation
Psoriasis1	Psoriasis (self-reported)
Psoriasis2	L40 Psoriasis
Thyroid1	Thyroid problem (not cancer; self-reported)
Thyroid2	E00-E07 Disorders of thyroid gland
UC	K51 Ulcerative colitis
VTE	Venous thromboembolic disease (self-reported)

935

936 **Figure 3: Co-localisation of SHPS1 (encoded by *SHPS1*: synonym *SIRPA*) and**
937 **schizophrenia DNA associations.** Upper panel, locuszoom (Pruim et al., 2010) of the region
938 surrounding *SHPS1* and the associations with schizophrenia (Schizophrenia Working Group
939 of the Psychiatric Genomics Consortium et al., 2014); lower panel, associations with SHPS1.
940 Lower panel inset, the relative concentration of SHPS1 across the 3 genotypes of rs4813319
941 – the DNA variant used as the instrumental variable (IV) in the MR analysis: CC, CT, and TT.

942

943 **Figure 3.**



944

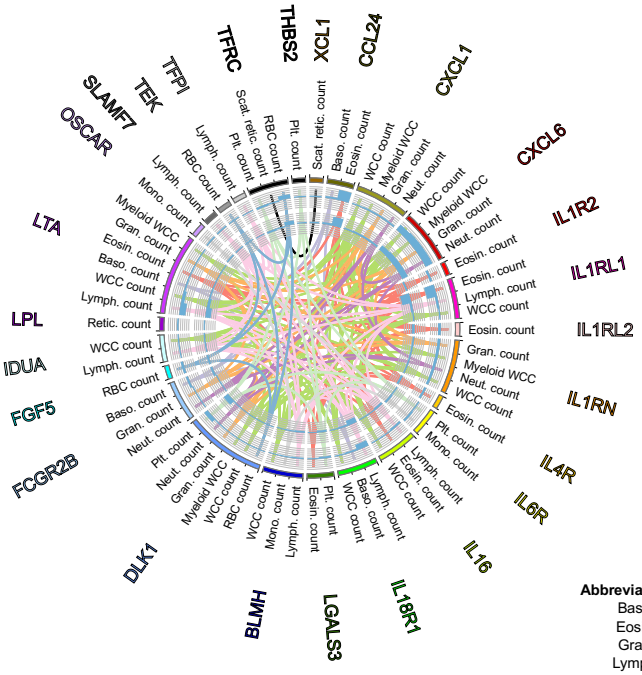
945 **Supplementary Figure S1**. Significant (FDR <0.05) P²MR protein-outcome causal
946 inferences: haematology count subset.

947 a) Phenoscanner (Staley et al., 2016): P²MR significant protein-haematology count outcome
948 causal inferences for 20 Phenoscanner studies. b) GeneAtlas (Canela-Xandri et al., 2017): MR
949 significant protein-haematology count outcome causal inferences for UK Biobank data.
950 Asterisks indicate P²MR estimates that are not significantly heterogeneous (HEIDI, Main
951 Text (Zhu et al., 2016)). Key as for Figure 2: Reading from the outside in: protein (exposure;
952 HGNC symbol); haematology count outcome; key colour; bar chart of the signed
953 (beta/standard error)² value of the MR estimate (using pQTL data from the discovery
954 cohort; Methods); and bar chart of the signed (beta/standard error)² value of the MR
955 estimate (using pQTL data from the replication cohort; Methods). Central chords join
956 identical outcomes. Identically coloured chords indicate similar outcome groups.

957

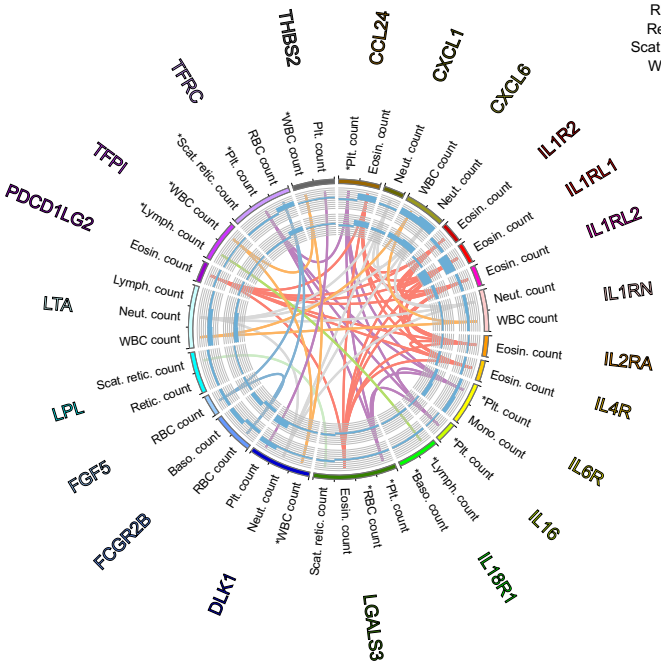
Supplementary Figure S1.

a) Phenoscanner



Abbreviated Outcome	Outcome Description
Baso. count	Basophil count
Eosin. count	Eosinophil count
Gran. count	Granulocyte count
Lymph. count	Lymphocyte count
Mono. count	Monocyte count
Myeloid WCC	Myeloid white cell count
Neut. count	Neutrophil count
Plt. count	Platelet count
RBC count	Red blood cell count
Retic. count	Reticulocyte count
Scat. retic. count	High light scatter reticulocyte count
WCC count	White blood cell count

b) UKBB

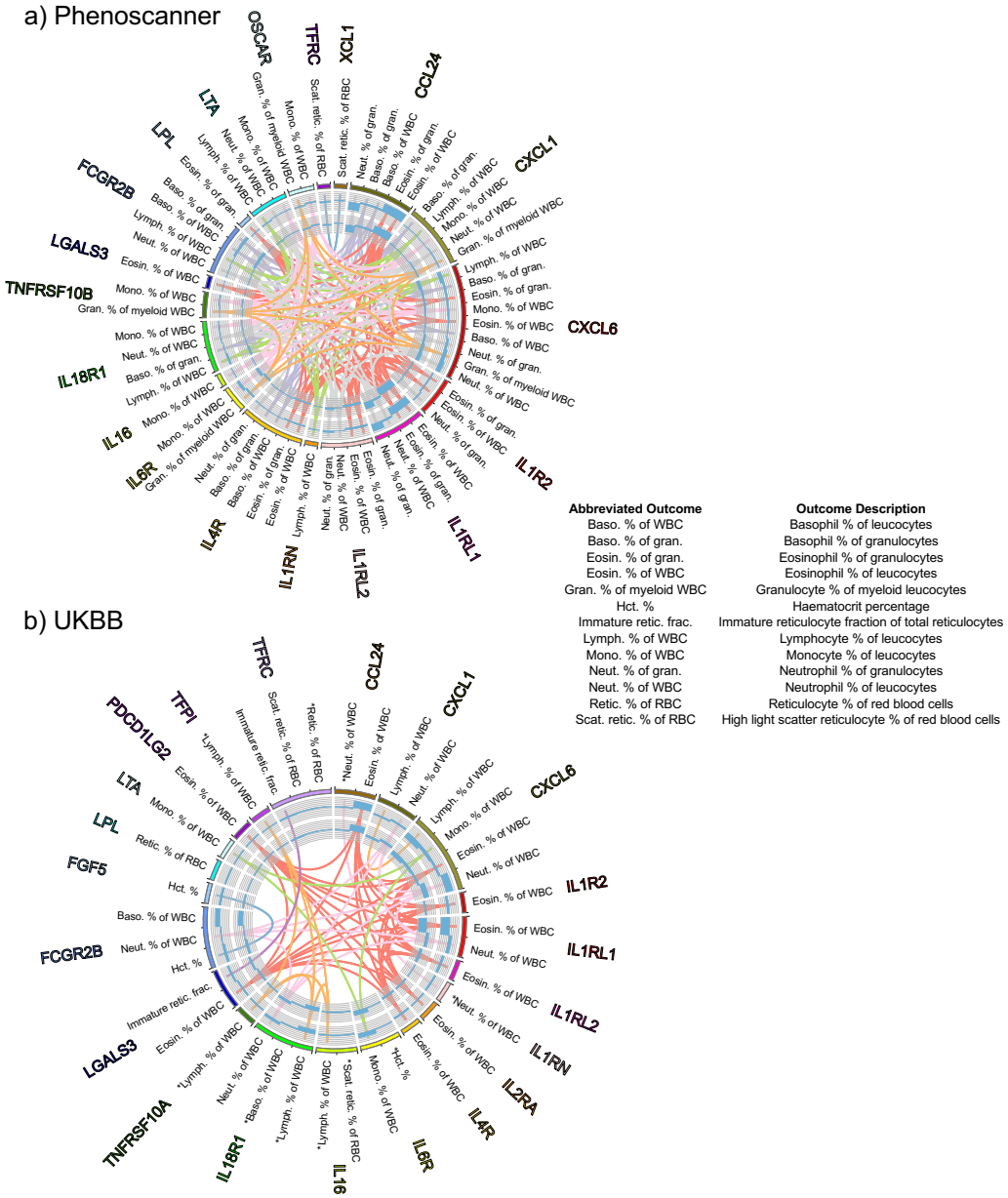


960 **Supplementary Figure S2**. Significant (FDR <0.05) P²MR protein-outcome causal
961 inferences: haematology percentage subset.

962 a) Phenoscanner (Staley et al., 2016): P²MR significant protein-haematology percentage
963 outcome causal inferences for 20 Phenoscanner studies. b) GeneAtlas (Canela-Xandri et al.,
964 2017): MR significant protein-haematology percentage outcome causal inferences for UK
965 Biobank data. Asterisks indicate P²MR estimates that are not significantly heterogeneous
966 (HEIDI, Main Text (Zhu et al., 2016)). Key as for Figure 2: Reading from the outside in:
967 protein (exposure; HGNC symbol); haematology percentage outcome; key colour; bar chart
968 of the signed (beta/standard error)² value of the MR estimate (using pQTL data from the
969 discovery cohort; Methods); and bar chart of the signed (beta/standard error)² value of the
970 MR estimate (using pQTL data from the replication cohort; Methods). Central chords join
971 identical outcomes. Identically coloured chords indicate similar outcome groups.

972

973 **Supplementary Figure S2.**



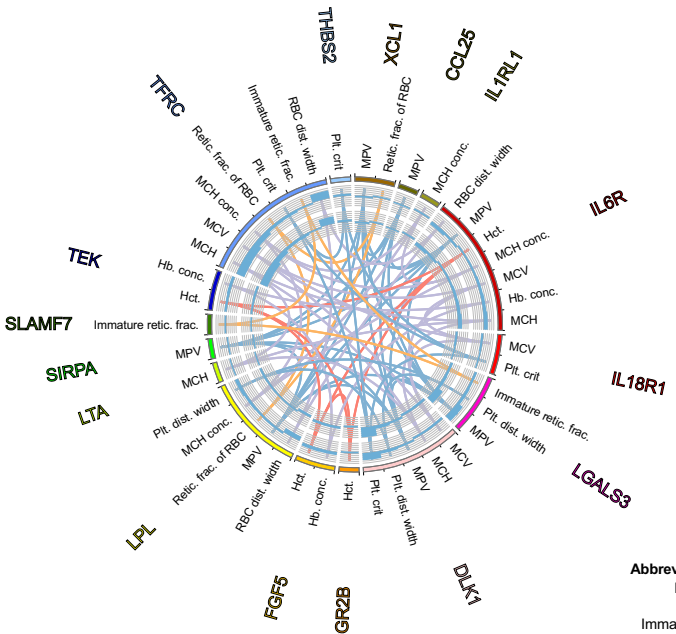
974 **Supplementary Figure S3**. Significant (FDR <0.05) P²MR protein-outcome causal
975 inferences: haematology (non-count, non-percentage) subset.

976 a) Phenoscanner (Staley et al., 2016): P²MR significant protein-haematology outcome causal
977 inferences for 20 Phenoscanner studies. b) GeneAtlas (Canela-Xandri et al., 2017): MR
978 significant protein-haematology outcome causal inferences for UK Biobank data. Asterisks
979 indicate P²MR estimates that are not significantly heterogeneous (HEIDI, Main Text (Zhu et
980 al., 2016)). Key as for Figure 2: Reading from the outside in: protein (exposure; HGNC
981 symbol); haematology outcome; key colour; bar chart of the signed (beta/standard error)²
982 value of the MR estimate (using pQTL data from the discovery cohort; Methods); and bar
983 chart of the signed (beta/standard error)² value of the MR estimate (using pQTL data from
984 the replication cohort; Methods). Central chords join identical outcomes. Identically
985 coloured chords indicate similar outcome groups.

986

Supplementary Figure S3.

a) Phenoscanner



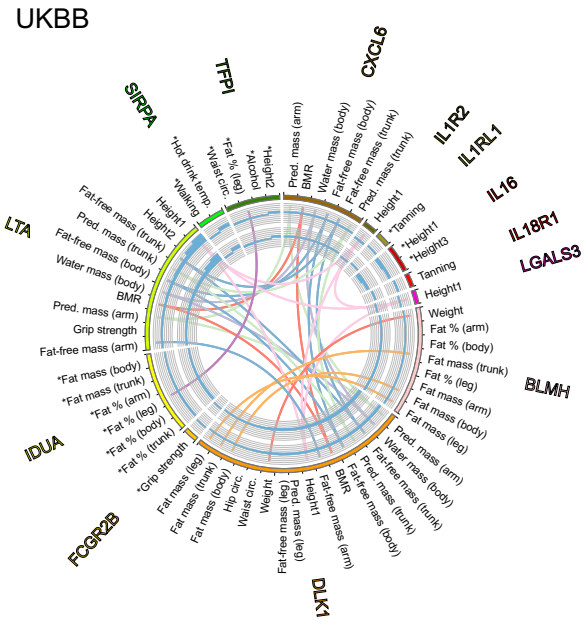
b) UKBB

Abbreviated Outcome	Outcome Description
Hb. conc.	Haemoglobin concentration
Hct.	Hematocrit
Immature retic. frac.	Immature frac. of reticulocytes
MCH	Mean corpuscular haemoglobin
MCH conc.	MCH concentration
MCV	Mean corpuscular volume
MPV	Mean platelet volume
MRV	Mean reticulocyte volume
MSV	Mean sphered cell volume
Plt. crit	Platelet crit
Plt. dist. width	Platelet distribution width
RBC dist. width	Red blood cell distribution width
Retic. frac. of RBC	Reticulocyte frac. of red cells

Supplementary Figure S4. Significant (FDR <0.05) P^{2}MR protein-outcome causal inferences: anthropometric measurements subset.

GeneAtlas (Canela-Xandri et al., 2017): MR significant protein-anthropometric measurements outcome causal inferences for UK Biobank data. Asterisks indicate P^{2}MR estimates that are not significantly heterogeneous (HEIDI, Main Text (Zhu et al., 2016)). Key as for Figure 2: Reading from the outside in: protein (exposure; HGNC symbol); anthropometric measurements outcome; key colour; bar chart of the signed (beta/standard error)^2 value of the MR estimate (using pQTL data from the discovery cohort; Methods); and bar chart of the signed (beta/standard error)^2 value of the MR estimate (using pQTL data from the replication cohort; Methods). Central chords join identical outcomes. Identically coloured chords indicate similar outcome groups.

1001 **Supplementary Figure S4.**



Abbreviated Outcome	Outcome Description
Alcohol	Alcohol intake frequency
BMR	Basal metabolic rate
Fat % (arm)	Arm fat percentage
Fat % (body)	Body fat percentage
Fat % (leg)	Leg fat percentage
Fat % (trunk)	Trunk fat percentage
Fat mass (arm)	Arm fat mass
Fat mass (body)	Whole body fat mass
Fat mass (leg)	Leg fat mass
Fat mass (trunk)	Trunk fat mass
Fat-free mass (arm)	Arm fat-free mass
Fat-free mass (body)	Whole body fat-free mass
Fat-free mass (leg)	Leg fat-free mass
Fat-free mass (trunk)	Trunk fat-free mass
Grip strength	Hand grip strength
Height1	Standing height
Height2	Sitting height
Height3	Comparative height size at age 10
Hip circ.	Hip circumference
Hot drink temp.	Hot drink temperature
Pred. mass (arm)	Arm predicted mass
Pred. mass (leg)	Leg predicted mass
Pred. mass (trunk)	Trunk predicted mass
Tanning	Ease of skin tanning
Waist circ.	Waist circ.
Waist circ. / Hip circ.	Waist circ. / Hip circ.
Walking	Number of days/week walked 10+ minutes
Water mass (body)	Whole body water mass
Weight	Weight

1002