bioRxiv preprint doi: https://doi.org/10.1101/854752; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Whole genome sequencing analysis of the cardiometabolic proteome

Arthur Gilly[1,2], Young-Chan Park[2,3], Grace Png[1,2], Thea Bjornland[2,4], Lorraine Southam[1,2,5], Daniel Suveges[2,6], Sonja Neumeyer[1], Iris Fischer[1], Andrei Barysenka[1], N. William Rayner[2,7,8], Emmanouil Tsafantakis[9], Maria Karaleftheri[10], George Dedoussis[11], Eleftheria Zeggini[1,2*]

1. Institute of Translational Genomics, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany
2. Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton CB10 1SA, UK
3. University of Cambridge, Cambridge, UK
4. Department of Mathematical Sciences, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway
5. Wellcome Centre for Human Genetics, Oxford, UK
6. European Bioinformatics Institute, Wellcome Genome Campus, Hinxton CB10 1SH, UK
7. Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, UK
8. Oxford Centre for Diabetes, Endocrinology and Metabolism, Radcliffe Department of Medicine, University of Oxford, Oxford, UK
9. Anogia Medical Centre, Anogia, Greece
10. Echinos Medical Centre, Echinos, Greece
11. Department of Nutrition and Dietetics, School of Health Science and Education, Harokopio University of Athens, Greece

*Correspondence to: eleftheria.zeggini@helmholtz-muenchen.de

bioRxiv preprint doi: https://doi.org/10.1101/854752; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The human proteome is a crucial intermediate between complex diseases and their genetic and environmental components, and an important source of drug development targets and biomarkers. Here, we conduct high-depth (22.5x) whole-genome sequencing (WGS) in 1,328 individuals to fully assess the genetic architecture of 257 circulating protein biomarkers of cardiometabolic relevance. We discover 132 independent sequence variant associations (P<7.45x10^{-11}) across the allele frequency spectrum, including 44 new *cis*-acting and 11 new *trans*-acting loci, all of which replicate in an independent cohort (n=1,605, 18.4x WGS). We identify replicating evidence for rare-variant *cis*-acting protein quantitative trait loci for five genes, involving both coding and non-coding variation. We find causal links between protein biomarkers and cardiovascular, inflammatory and immune-related diseases. We construct and validate polygenic risk scores that explain up to 45% of protein level variation, and find significant correlation between genetically-predicted biomarker levels and cardiovascular disease risk in UK Biobank.

Cardiometabolic diseases are a leading cause of death and continue to rise in prevalence across global populations. Biomarkers can help improve early detection and diagnosis, and can lead to better clinical outcomes through improved and timely interventions. To comprehensively characterise the genetic underpinning of protein biomarker levels, disentangle correlation from causation, and identify opportunities for new therapeutic target discovery and predictive modeling, we first assessed the association between 257 cardiometabolic disease-related serum protein levels[1] and 13,419,876 single nucleotide variants (SNVs) in a population-based cohort (MANOLIS) with deep whole genome sequence data.

We identify 116 protein quantitative trait loci (pQTLs) reaching study-wide significance (P<7.45x10^{-11}) (Supplementary Table 1, Figure 1). Thirty-two (27%) of these are driven by multiple independent variants (between two and seven per locus; for a total of 164 independently-associated variants) illustrating complex allelic architecture at pQTLs (Supplementary Figure 1). We find replicating evidence for association (P<0.000305) across 131 out of 159 variants (82%) present in an independent, whole genome-sequenced population-based cohort with the same serum biomarker measurements (Pomak[2]) (n=1,605, 18.4x WGS, Supplementary Figure 2). We find that these robustly-replicating loci explain up to 47.7% of protein level variance, and on average more (one-sided Mann-Whitney-U test, P=3.42x10^{-13}) than for 37 other, non-proteomic quantitative traits measured in the same individuals (Supplementary Figure 3). This exemplifies how the study of blood biomarkers can powerfully capture the heritable component of biological processes underpinning such disease-relevant quantitative traits.

117 (90%) of these 131 reproducibly-associated variants are common (minor allele frequency (MAF) >5%), 13 are low-frequency (MAF 1-5%) and 1 is rare (MAF<1%) (Figure 2). One hundred of the associated variants (76%) are located within 1Mb of the gene encoding the respective protein (i.e. in *cis*-pQTLs), and 31 (24%) are in *trans*-pQTLs (Figure 2). 57 *cis*-associated variants are located within the boundaries of the respective gene, and the remaining 43 are at a median distance of 13.8kb (max. 920 kb) (Supplementary Figure 4). The majority (48/72, 66%) of *cis*-pQTLs discovered in this cohort have either not previously been reported in large association studies[3-7], or harbour variants conditionally independent of all previously-reported associations (Supplementary Table 1).

bioRxiv preprint doi: https://doi.org/10.1101/854752; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We identify 38 variants in 35 *trans* loci associated with 32 proteins; 22 of these variants have not been previously reported (Supplementary Text) in large protein-level GWAS. We find the replication rate in Pomak to be similar for *trans*- (81%) and *cis*-associated (79%) variants. Of the replicating 31 *trans*-acting variants, 30 are common and one is low-frequency. We identify *trans*-pQTL signals for seven receptor/ligand pairs with experimental evidence of physical interaction and well-established synergistic roles in downstream pathways[8-14] (Table 1).

To enhance our understanding of rare variant (RV) contribution to serum protein biomarker levels, we performed gene-based burden analysis across coding and non-coding sequence variation (Methods). We identify 6 study-wide significant ($P<7.45\times10^{-11}$) *cis*-RV-pQTLs (Figure 3, Supplementary Figure 5), in the *ACP6* (lysophosphatidic acid phosphatase type 6, $P_{meta\text{-}analysis}=3.17\times10^{-97}$) , *PON3* (paraoxonase 3, $P_{meta\text{-}analysis}=7.42\times10^{-86}$), *IL1RL1* (interleukin 1 receptor like 1, $P_{meta\text{-}analysis}=2.15\times10^{-58}$), *DPP7* (dipeptidyl peptidase 7, $P_{meta\text{-}analysis}=2.71\times10^{-36}$), *CTSO* (cathepsin O, $P_{meta\text{-}analysis}=2.27\times10^{-33}$) and *GRN* (progranulin, $P_{MANOLIS}=3.16\times10^{-12}$) genes. All except the *GRN* burden signal replicate in the Pomak cohort. The *GRN* RV-pQTL is driven by the novel splice donor variant chr17:44349552 (minor allele count MAC=4) and the 5'UTR variant rs563336550 (MAF=1.7%) in MANOLIS, the latter showing a 17-fold increase in frequency in MANOLIS compared to gnomAD non-Finnish Europeans (MAF=0.1%), and more than 2000-fold compared to TOPMed (MAF=0.00079%).

We find that rare regulatory variants are major contributors to some of these burdens. For example, one of the two variants driving the PON3 *cis*-RV-pQTL resides in promoter ENSR00000215353 and transcription factor binding site ENSR00000832511, and is associated with a decrease in PON3 levels (rs149867961, MAF=3.1%, effect size β=-1.18 in units of standard deviation, standard error σ=0.113, $P=7.58\times10^{-23}$). The other contributing variant, rs772677677 (MAF=1.9%, β=-1.55, σ=0.143, $P=3.15\times10^{-24}$), is a missense variant with a substantially increased frequency in MANOLIS (MAF=1.9% compared to 0.00264% in gnomAD), also associated with a decrease in PON3 levels. PON3 (paraoxonase 3) inhibits the oxidation of low-density lipoprotein (LDL), an effect that slows atherosclerosis progression[15]. These findings illustrate the contribution of rare variants to the heritability of proteomic traits, and that this contribution is partly mediated through *cis*-RV-pQTLs.

To detect proteins that may play a causal role in cardiometabolic disease onset or progression, we performed two-sample Mendelian randomization analysis across 93 proteins with study-wide significant signals in our study and 193 diseases and traits from UK Biobank and other large consortial datasets (Methods, Supplementary Table 2). We identify significant (FDR<0.05) associations involving 48 proteins and 75 phenotypes (Supplementary Table 3, Figure 4).

Providing proof of principle, we find evidence of established links between dysregulated protein levels and common diseases, such as an inverse causal correlation between PCSK9 levels and hypercholesterolemia ($P=1.00\times10^{-10}$, $P_{FDR}=9.46\times10^{-8}$), and between osteopontin levels and risk of osteoporosis ($P=6.49\times10^{-5}$, $P_{FDR}=0.011$) and hypothyroidism ($P=1.2\times10^{-4}$, $P_{FDR}=0.016$). Similarly, we find decreased levels of IL1RL1 and IL1RT2, both proteins involved in autoimmunity and inflammation-related disorders[16], to be causally linked to risk of autoinflammatory bowel diseases (Supplementary Table 3).

bioRxiv preprint doi: https://doi.org/10.1101/854752; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We further identify new evidence for disease-mediating roles for proteins circulating in the periphery. For example, rs2306272, a missense *cis*-pQTL, is associated with decreased LRIG1 (leucine rich repeats and immunoglobulin like domains 1) levels (MAF=31%, meta-analysis β=-0.754, σ=0.0261, P=1.50x10^{-183}), and is causally associated with reduced risk of cardiometabolic diseases, including atrial fibrillation (P=5.23x10^{-11}, P_{FDR}=5.32x10^{-8}) and type 2 diabetes (P=4.70x10^{-5}, P_{FDR}=8.87x10^{-3}), and with lower BMI (P=2.72x10^{-9}, P_{FDR}=1.89x10^{-6}) (Figure 4). Variants in *LRIG1* have previously been associated with atrial fibrillation[17,18], and the identified pQTL co-localises with previous pulse rate (posterior probability of colocalisation P_4=0.939) and QRS duration (P_4=0.998) loci. Mouse knockout models of *LRIG1* exhibit decreased body weight and fat[19].

Notably, we find evidence for a genetic link between *PRG2* intronic variant rs10642232 and decreased levels of PAPPA (pregnancy-associated plasma protein-A) (β=-0.299, σ=0.0320, P=1.06x10^{-20}). PAPPA is a metalloproteinase involved in normal and pathological insulin-like growth factor (IGF) physiology. *PRG2* codes for eosinophil granule major basic protein, which reduces PAPPA activity by interacting with it to form a complex[20]. PAPPA is a specific protease targeting IGFBP4 (IGF binding protein 4) in the presence of IGF. IGFBP4 inhibits IGF binding with its receptor, and PAPPA promotes IGF activity[21]. rs10642232 is a *PRG2*-decreasing eQTL in multiple tissues. We find reduced PAPPA levels to be causally associated with decreased risk of diabetic kidney disease in T2D patients (P=2.63x10^{-4}, P_{FDR}=0.0304). IGF activity is enhanced in early diabetic nephropathy, whereas IGF resistance is found in chronic kidney failure[22]. Animal knockouts of PAPPA exhibit decreased body weight and length, type 2 diabetes and hypercholesterolemia. Our results suggest that PAPPA and its inhibition by PRG2 within the IGF system may play a role in the pathogenesis and progression of diabetic kidney disease in T2D patients. These findings are consistent with the reported lower incidence of diabetic complications in this population.

Further, we find that *cis*-acting variants decreasing levels of ENTPD5 (rs73301485, MAF=7.5%, β=-0.637, σ=0.0487, P=4.59x10^{-39}; rs140111715, MAF=3.7%, β=-0.778, σ=0.0563, P=2.29x10^{-43}) are causally associated with lower risk of type 2 diabetes (P=3.56x10^{-4}, P_{FDR}=0.037) and diabetic kidney disease (P=2.61x10^{-19}, P_{FDR}=4.93x10^{-16}). ENTPD5 (ectonucleoside triphosphate diphosphohydrolase 5) is primarily expressed in the liver, kidney, intestine, prostate and bladder, and rs73301485 is an eQTL in multiple tissues. Mouse knockout models show decreased body weight, hypoglycemia, decreased cholesterol and triglycerides[23]. Small-molecule screens have recently identified several ENTPD5 inhibitors[24] that warrant investigation for their effect on type 2 diabetes and diabetic complications.

We find that genome-wide polygenic risk scores (PRS) calculated in MANOLIS can predict up to 45.5% of protein variance in the independent Pomak dataset, despite a low average predictive performance (median r^2=0.026, Figure 5). We find allele frequency thresholds not to have an appreciable influence on predictive power. The scores that achieve high accuracy (r^2>0.05) in predicting Pomak protein levels all involve stringent thresholds (P<1x10^{-6}, with the majority at P<1x10^{-9}, Supplementary Table 4). This indicates an architecture involving a small number of strongly-associated common variants, and a smaller contribution for rare and low-frequency variants, within the power constraints of the study. Notably, both the discovery and test datasets stem from individuals of European ancestry and further studies in global populations will be required to assess the transferability of these PRS.

bioRxiv preprint doi: https://doi.org/10.1101/854752; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Polygenic prediction of the cardiometabolic proteome can lead to the identification of potential biomarkers through correlation with disease states. We performed logistic regression of 47 proteins with PRS that had achieved a predictive value of r^2>0.05 in Pomak on 86 indications in UK Biobank, adjusted for genetic principal components, clinical and lifestyle factors. We find that PRS for GRN (progranulin), CHI3L1 (chitinase 3 like 1) and PECAM1 (platelet and endothelial cell adhesion molecule 1) levels are significant predictors of disease status (Wald test $P<1.66\times10^{-5}$) across a range of cardiometabolic traits (Table 2). The progranulin level PRS is correlated with increased risk of hypercholesterolemia, and is driven by a single association in *CELSR2-SORT1*, an established risk locus for lipid disorders[25]. In a joint predictive model for high cholesterol, inclusion of PRS for CHI3L1 and PECAM1 levels results in a significant increase of the model accuracy compared to the clinical and lifestyle covariates-only model (Supplementary Text, Likelihood Ratio Test $P=7.24\times10^{-27}$). We find that an elastic net model agnostically selects the same three PRS in a full-proteome analysis of high-cholesterol, confirming their contribution and demonstrating the value of including proteomics PRS in predictive models of disease risk.

In summary, using whole-genome sequencing, we identify robustly-replicating *cis*- and *trans*-pQTLs, and show for the first time that burdens of rare variants contribute to the genetic architecture of protein biomarker levels. We show that incorporating information on this genetic contribution leads to improvement in clinical risk models for cardiovascular disease. Identification of causal contributions of the cardiometabolic proteome to the risk of multiple chronic diseases can present opportunities for new therapeutic target discovery and predictive modeling to accelerate precision medicine.

bioRxiv preprint doi: https://doi.org/10.1101/854752; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Methods

Sequencing and variant calling

Genomic DNA (500 ng) from 1482 samples was subjected to standard Illumina paired-end DNA library construction. Libraries underwent DNA sequencing using the HiSeqX platform (Illumina) according to manufacturer's instructions. Variant calling was performed after filtering out contamination, spatial artifacts and duplicates, according to the Genome Analysis Toolkit (GATK) v. 3.5-0-g36282e4 Best Practices. The sequencing and variant calling is described in detail in a previous publication[26].

Proteomics

The serum levels of 276 unique proteins in 1,407 MANOLIS samples from three Olink panels - CVDII, CVDIII and Metabolism - were measured using Olink's proximity extension assay (PEA) technology[1]. Briefly, for each assay, the binding of a unique pair of oligonucleotide-labelled antibody probes to the protein of interest results in the hybridisation of the complementary oligonucleotides, which triggers extension of by DNA polymerase. DNA barcodes unique to each protein are then amplified and quantified using microfluidic real-time qPCR. Measurements were given in a natural logarithmic scale in Normalised Protein eXpression (NPX) levels, a relative quantification unit. NPX is derived by first adjusting the qPCR Ct values by an extension control, followed by an inter-plate control and a correction factor predetermined by a negative control signal. This is followed by intensity normalisation, where values for each assay are centered around its median across plates to adjust for inter-plate technical variation. Further details on the internal and external controls used can be found at http://www.olink.com. Additionally, a lower limit of detection (LOD) value is determined for each protein based on the negative control signal plus three standard deviations. For our samples, NPX values that fall below the LOD were set to missing.

We adjusted all phenotypes using a linear regression for age, age squared, sex, plate number, and per-sample mean NPX value across all assays, followed by normalisation of the residuals. We also adjusted for season, given the observed annual variability of some circulating protein levels. Given the dry Mediterranean climate of Crete, we define season of collection as hot summer or mild winter. Plate effects are partially offset by the median-centering implemented by Olink. MANOLIS samples were plated in the order of sample collection, which results in plate and season information to be largely correlated.

Quality control

We excluded 18 proteins across all panels with missingness or below-LOD proportion greater than 40%. BNP was measured across all three panels, and was excluded due to high missingness in CVDII. In total, we therefore excluded 19 protein measurements (total analysed 257, Supplementary Table 5). 26, 2 and 14 samples failed vendor QC and were excluded from CVDII, III and META, respectively. 42 samples were excluded due to missing age.

Sequencing data quality control has been described before[26]. Briefly, Variant-level QC was performed using the Variant Quality Score Recalibration tool (VQSR) from the Genome Analysis Toolkit (GATK) v. 3.5-0-g36282e4. Sample-level QC was performed by comparing genotypes with chip data in the same samples. Four individuals failed sex checks, 8 samples had low concordance with chip data, 11 samples were duplicates, and 12 samples displayed traces of contamination. As contamination and

bioRxiv preprint doi: https://doi.org/10.1101/854752; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

sex mismatches were correlated, a total of 25 individuals were excluded (n = 1457). Variants were further filtered using the Hardy-Weinberg equilibrium test at $P = 1.0 \times 10^{-5}$. We filtered out 14% of variants with call rates < 99%.

Single-point association

We carry out single-point association using the linear mixed model implemented in GEMMA [27]. We use an empirical relatedness matrix calculated on a LD-pruned set of low-frequency and common variants (MAF>1%) that pass the Hardy-Weinberg equilibrium test (P<1x10^{-5}). We further filter out variants with missingness higher than 1% and MAC <10. 5 proteins were excluded due to having a genomic control λ_{GC}<0.97 orλ_{GC}>1.05 after association (Supplementary Table 5). 123 signals were extracted using the PeakPlotter software (https://github.com/hmgu-itg/peakplotter), which is based on a combination of distance-based and LD-based pruning. We extracted independent SNV at each associated locus using an approximate conditional and joint stepwise model selection analysis as implemented in GCTA-COJO[28]. To avoid overfitting when too many predictors are included in the model, we perform LD-based clumping using Plink v.1.9, based on an r^2 value of 0.1 and a window of 1Mb prior to the GCTA-COJO analysis[29]. The extended linkage disequilibrium (LD) present within MANOLIS can cause very large peaks to be broken up into several signals. We identified and manually investigated 5 regions where multiple peaks were present, reducing the number of independent signals to 116 and the number of conditionally independent variants to 164.

Rare variant association

For rare variant association, we apply a MAF filter of 5% and a missingness filter of 1%. We use the linear mixed model extension of SKAT-O implemented in MONSTER[30], using the MUMMY wrapper[26] (https://github.com/hmgu-itg/burden_testing). Following our previously-reported analysis strategy[26], we test for rare variant burden association on a gene-by-gene basis: firstly, restricting burdens to coding variants with Ensembl most severe consequence stronger than missense; secondly, including all coding variants weighted by CADD[31]; thirdly, including exon and regulatory variants using the phred-scaled Eigen score[32]; and, finally, regulatory variants only weighted by Eigen. Regulatory regions are linked to a gene if they overlap an Ensembl-documented eQTL for that gene in any tissue. A gene-pair was taken forward for quality control (QC) if it was significant in any of the four analyses. 17 signals passed this threshold. Because rare variant LD blocks can extend over long distances and capture overlapping common associations, we manually inspect LD blocks through the plotburden software (https://github.com/hmgu-itg/plotburden), and discard signals involving variants in LD with nearby *cis* ones. For each remaining signal, we then re-run the burden analysis conditional on the genotypes of the variant with the lowest single-point p-value that was previously included in the burden, so as to only consider signals arising from at least two distinct variants. 6 RV-pQTL signals pass this quality control procedure (Supplementary Figure 5).

Replication

We performed replication in 1,605 samples from the Pomak cohort sequenced at a mean depth of 18.6x, using an identical sequencing, variant calling and quality control protocol. The proteomic phenotypes were transformed identically to MANOLIS. Single-point association analysis was performed using GEMMA and an identically calculated GRM. For burden replication, we specifically analysed the genes associated in MANOLIS in all conditions, and defined replication if a significant signal was detected in any of them. The COMT signal driven by rs4680 in MANOLIS (β=-0.373,

bioRxiv preprint doi: https://doi.org/10.1101/854752; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

σ=0.0405, P=3.5x10^{-20}) could not be replicated due to COMT failing QC in Pomak. In five cases, associated MANOLIS variants and those tagged (r^2>0.8) by them were monomorphic in the Pomak cohort (rs183455943, rs4778724, rs1053361963, rs200251994, rs186044494). 101/116 (87%) loci had at least one replicating variant.

Definition of novelty

To assess whether a protein had been previously studied, we examined protein lists and summary statistics from five large published proteomics GWAS[3-7]. In total, 44 proteins measured in MANOLIS did not match proteins previously assayed in these studies. To determine novelty of genetic *cis* and *trans* association with proteins in our study, we first determined previously reported variants within a 1Mb window around the association peaks. We then used GEMMA[27] to perform association analysis using known variants as covariates. The association signals were declared novel if either there were no known signals in the vicinity, or the associations were still significant (P-value threshold: 1.00x10^{-5}) after conditioning. For *trans* associations, we further annotated signals depending on whether they fell within highly pleiotropic genes (*KLKB1, ABO, APOE, FUT2, F12*), whether they were independent of any *cis* signals in the vicinity, and whether the region or gene presented with any prior association in the Ensembl database. After this procedure, 58 *cis*-associated variants in 44 loci were either not within 1Mb or independent of a signal reported in previous large proteomics GWAS. 22 *trans*-associated variants were both novel and independent from *cis* loci. 11 of these were not located within highly pleiotropic genes.

Variant consequences

Consequence was evaluated using Ensembl VEP[33] for each variant with respect to any transcript of the *cis* gene for *cis*-associated variants and to the mapped gene for *trans*-associated variants. For *trans* associations, variants were manually mapped to any gene in a 1Mb window coding for known ligands or interactants when they were not contained within gene boundaries, as was the case for CXCL16 and LDLR. 16 replicating independent variants had a most severe consequence equal to or more severe than missense according to Ensembl VEP. For every variant, we extracted tagging SNVs at r^2>0.8 using PLINK, however none of these tagging variants had a more severe consequence on the target gene than the independent variant. Similarly, we overlapped all independent variants with regulatory features using the Ensembl REST API. 35 variants in 29 loci overlapped with a regulatory feature. When extending the same analysis to variants with r^2>0.8 with independent variants, 93 variants in 68 loci could be mapped to a regulatory feature.

Colocalisation testing for eQTL overlap and pheWAS analysis

We perform colocalisation testing with eQTL data from the GTEX database[34]. First, for every signal, regions are extended 1Mb either side of every independent variant, and associations are conditioned on every other variant in the peak using GEMMA. For *cis* signals, expression information for the *cis* gene is extracted from the GTEX database over the same region. For *trans* signals, expression information is restricted to all genes located within a 2Mb region surrounding the variant. Then, for every variant/gene pair, we perform colocalisation testing using the fast.coloc function from the gtx R package (https://github.com/tobyjohnson/gtx). To account for multiple independent variants at the same locus, we perform conditional analysis on all independent variants except the one under consideration, and use these results as input for the colocalisation analysis. 60 (49%) independent variants in 45 (57%) *cis*-pQTLs co-localise with an expression quantitative trait locus (eQTL; identified

bioRxiv preprint doi: https://doi.org/10.1101/854752; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

in GTEx[35]) for the cis gene, in keeping with previous estimates for the proportion of pQTLs exerting their effects through transcriptional mechanisms[6]. In addition, we find that 29 (70%) *trans* signals co-localise with eQTLs for at least one gene in their vicinity (+/- 1Mb).

We apply the same procedure to PhenoScanner data[36], where eQTL data is replaced by the output of the PhenoScanner python utility 500kb either side of every independent variant. The colocalisation procedure is then repeated for every phenotype where at least one association is present in the region. Because such pheWAS results do not necessarily report weakly associated SNVs, the number of pheWAS variants to be colocalised can be small. We therefore modify the colocalisation script to handle the case where only one variant is present per phenotype. 82% of all signals co-localised with previous phenotypic associations (Supplementary Table 6). The proportion was similar for *trans* and *cis* loci, with 30 out of 37 co-localising with previous association signals in the latter case. As a proof of concept, we recapitulate strong evidence for colocalization of a progranulin-lowering trans-pQTL in the *SORT1-CELSR2* locus with increased risk of elevated cholesterol, angina pectoris, ischaemic heart disease and other cardio-damaging phenotypes. *trans*-pQTL for all the six proteins associated with an ABO region variant also colocalise with a range of blood and cardiovascular phenotypes, in line with a large body of evidence linking the ABO blood group with multiple diseases.

Drug Target evaluation

For evaluating whether associated genes were drug targets, we used the OpenTargets[37] and DrugBank[38] databases. We accessed OpenTargets using the OpenTarget API. We converted the DrugBank XML file to flat files using the dbparser R package, and performed gene name matching using the USCS Gene Info database, downloaded May 6, 2019. 8 of the proteins for which a signal was detected at study-wide significance were targeted by drugs according to OpenTargets. This was true for 39 proteins when queried against the DrugBank database (Supplementary Table 7).

Mouse models

We use the Ensembl REST API to extract mouse orthologs for all of the 109 genes whose protein for which genetic associations were found in our study. According to the IMPC API, KO experiments were reported for 26 of these orthologs, 18 of these having phenotypes associated with a p-value smaller than $1x10^{-4}$ (Supplementary Table 8).

Two-Sample MR

We extracted variants characterized as independent signals by GCTA-COJO on a protein-by-protein basis across all *cis* and *trans* loci, and excluded novel variants without an rs-ID. For each remaining variant, we then considered summary statistics for all tagging positions with r^2>0.8, merged the resulting data frame with the exposure dataset by rs-ID. All such records originating from all independent signals were then merged by protein and carried over to MR analysis using the MRBase R package. We excluded *trans* pleiotropic loci (*ABO, KLKB1, FUT2, APOE, F12*). MR was performed on a set of 127 cardiometabolic traits available in the MRBase database (Supplementary Table 2). First, we selected self-reported disease codes from UK Biobank that were both of cardiometabolic relevance and had more than 999 cases in the second data release. We then performed the same search on primary and secondary ICD codes. Finally, we included case/control studies available in MR Base that had a sample size of at least 20,000, excluding both the Neale and Elsworth Biobank GWAS datasets, as well as studies categorized in the anthropometric, reproductive aging, personality,

bioRxiv preprint doi: https://doi.org/10.1101/854752; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

haemotological(sic), psychiatric/neurological, education, behavioural, sleeping and aging. We also manually excluded cancer phenotypes, and included "Inflammatory bowel disease" (ID 292) which was wrongly categorized as quantitative. Since all of our instruments involved small number of variants (≤10), we use the inverse-variant weighted method, except for single-instrument analyses where we use the Wald ratio which consists of dividing the instrument-outcome by the instrument-exposure regression coefficient.

We also leveraged summary statistics manually downloaded from recent large association studies for Chronic Kidney Disease[39], blood lipids[40], Atrial Fibrillation, Type-II Diabetes, Coronary Artery Disease, estimated glomerular filtration rate[41], albuminuria[42], and anthropometric traits[43].
Two proteins, PDL2 and TNFRSF10C, had both *cis* and *trans* associations where the *trans* did not fall in a highly pleiotropic gene. We performed 2-sample MR excluding *trans,* and found that no *cis* variant in TNFRSF10C could be found in selected external studies. This was due to the lead variant, 8:23108277 T/TG, being novel, and secondary variants, such as rs779159813, being very rare (MAF= 0.573%). For PDL2, the known *cis* signal driven by rs62556120 (β=0.416, σ=0.0276, P=2.92x10^{-51}) was causally associated with increased risk of ulcerative colitis and inflammatory bowel disease (Supplementary Table 9), whereas in the *cis+trans* analysis, the addition of rs10935473 attenuates that signal, but drives a causal association with height.

Significance thresholds

We calculate the significance thresholds by computing the effective number of variants, traits and analyses for every analysis requiring multiple-testing correction.

Single-variant analyses

The effective number of proteins was computed using the ratio of the eigenvalue variance to its maximum[44,45]:

$$M_{eff} = M\left(1 - (M-1)V_{\lambda_{obs}}/M^2\right) = 1 + \frac{tr(\Sigma^T \Sigma)}{M}$$

where $V_{\lambda_{obs}}$ is the variance of the eigenvalues of the correlation matrix. For the $M = 257$ Olink phenotypes in this study, $M_{eff} = 131.5$, which we round to 132. The resulting p-value threshold is 7.45x10^{-11}.

Rare-variant analyses

We report both single-point and rare variant burden signals, therefore increasing the multiple testing burden. The exact magnitude of this phenomenon being unknown, we performed a simulation study to compute the effective number of tests in case single-variant, variant-aggregation or both are reported in an association study. We find that reporting rare variant signals in combination with single-point signals at 5% and 1% MAF thresholds increased the multiple testing burden only marginally and by less than one order of magnitude (Supplementary Figure 6).

Two-sample MR

To correct for multiple testing in our MR analysis, we adjust p-values using FDR correction and examine significant results at an FDR of 0.05.

bioRxiv preprint doi: https://doi.org/10.1101/854752; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Polygenic prediction

To examine predictive accuracy, we compute polygenic risk scores using PRSice 2[46] with the Pomak as a target dataset. We assess scores between $1x10^{-4}$ and study-wide significance with intervals of $1x10^{-10}$. We further apply three MAF thresholds, at 0.05, 0.01 and MAC=10, which produces three best scores per protein. To evaluate the predictive power of proteins for complex disease, we compute all scores between $1x10^{-6}$ and study-wide significance with the same interval. We then run a logistic regression of 86 UK Biobank ICD codes and self-reported traits (Supplementary Table 10) on all such scores on a protein-by-protein basis for 47 proteins that achieved r^2>0.5 in the previous step. Sex, age, Qualification, Smoking status and BMI as well as 10 principal components are also added to the model. In order to accurately assess effect sizes, we select the most predictive score for each protein where at least one score threshold meets the Bonferroni-corrected P-value threshold (47 proteins, 64 effective phenotypes, $P<1.66x10^{-5}$) in the Wald test for variable contribution. We then run the same model as before, with only this score and covariates as predictors. We removed associations between 3 proteins (CTRC, ICAM2, SELE) and deep vein thrombosis and/or pulmonary embolism due to the association being entirely driven by *ABO* variants.

In order to validate our approach for variable selection, we pooled all scores for all proteins in a single model of self-reported high cholesterol, and used elastic net regression to shrink coefficients for non-informative predictors. We run 10-fold cross-validation with a 20% hold-out sample to optimize the lambda value, and run 11 models using different values of alpha, from 0 (ridge regression) to 1 (lasso) at intervals of 0.1. We use a value of lambda one standard deviation away from the minimal value to increase shrinkage. Overall, the performance was comparable across models, but small values of alpha yielded a better predictive performance on the holdout set. An alpha value of 0.1 yielded almost no loss of AUC compared to ridge regression, but shrunk almost all coefficients to 0 (Supplementary Figure 7). Four protein scores (CHI3L1, PECAM1, SELE and GRN) were shrunk to a value greater than 0, confirming results of the manual variable selection procedure.

Acknowledgments

We thank the residents of the Pomak and Mylopotamos villages for taking part. The MANOLIS study is dedicated to the memory of Manolis Giannakakis, 1978–2010. This work was funded by the Wellcome Trust [098051] and the European Research Council [ERC-2011-StG 280559-SEPI]. The GATK3 program was made available through the generosity of the Medical and Population Genetics program at the Broad Institute, Inc. We thank the Human Genetics DNA Pipelines and Human Genetics Informatics departments at the Wellcome Sanger Institute for performing sequencing and variant calling. This study has been conducted using the UK Biobank Resource (project ID 10205).

Author Contributions

Sample collection and phenotyping: ET, MK, GD, EZ
Phenotype transformation and quality control: AG, YCP, GP, LS, NWR
Association analysis: AG, YCP
Software development: AG, DS
Bioinformatics: AG, SN, IF, AB
Simulation analysis: TB
Study design and supervision: EZ
Manuscript writing: AG, IF, EZ

bioRxiv preprint doi: https://doi.org/10.1101/854752; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Competing Interests statement

The authors declare no competing interests.

Tables

									Discovery (MANOLIS)				Replication (Pomak)				Meta-analysis			
protein	**rsID**	**chromosome**	**position**	**gene**	**consequence**	**previously known**	**EA**	**NEA**	**MAF (%)**	**β**	**σ**	**Score test *P***	**MAF (%)**	**β**	**σ**	**Score test *P***	**β**	**σ**	**Inverse-variance weighted *P***	**I^2**
CXCL1	rs12075	1	159205564	*ACKR1*	missense	no	G	A	49.3	-0.413	0.0381	$2.81x10^{-27}$	35.4	-0.667	0.0337	$2.61x10^{-63}$	-0.556	0.0252	$2.12x10^{-107}$	0.96
MCP1	rs12075	1	159205564	*ACKR1*	missense	yes[6]	G	A	49.5	-0.585	0.0362	$7.87x10^{-59}$	35.3	-0.776	0.0321	$8.50x10^{-84}$	-0.692	0.0240	$8.33x10^{-183}$	0.936
CXCL16	novel	3	45948068	*CXCR6*	downstream	no	T	C	0.9	1.51	0.215	$8.46x10^{-12}$	0 (MAC=1)	-1.85	0.967	$5.60x10^{-02}$	1.35	0.210	$1.15x10^{-10}$	0.913
TNFRSF13B	rs200748895	13	108308032	*TNFSF13B*	3' UTR	no	T	TGCTG	7.4	0.508	0.0731	$3.83x10^{-12}$	4.8	0.670	0.0848	$1.66x10^{-14}$	0.577	0.0554	$2.08x10^{-25}$	0.522
UPA	rs4251805	19	43670289	*PLAUR*	5' UTR	yes[6]	T	C	5.0	-0.594	0.0862	$5.52x10^{-12}$	3.3	-0.536	0.104	$3.57x10^{-07}$	-0.570	0.0664	$8.45x10^{-18}$	0
LDLR	rs1065853	19	44909976	*APOE*	downstream	no	T	G	7.7	0.507	0.0724	$2.45x10^{-12}$	8.4	0.527	0.0624	$3.52x10^{-16}$	0.519	0.0472	$5.06x10^{-28}$	0
IL18	rs385076	2	32264782	*NLRC4*	5’ UTR	yes[47,48]	C	T	34.9	-0.281	0.0415	$1.27x10^{-11}$	42.5	-0.324	0.0371	$5.63x10^{-17}$	-0.305	0.0277	$3.10x10^{-28}$	0

Table 1: Receptor-ligand *trans*-pQTLs. All positions are on genomic build 38. Effect alleles are always the minor allele, and are all alternate with respect to the reference, except for the CXCL1, MCP1 and IL18 variants. SE: standard error, MAF: minor allele frequency, MAC: minor allele count, EA: effect allele, NEA: non-effect allele, I^2: I-squared statistic of study heterogeneity[49,50]. Consequences were calculated for the mapped gene using Ensembl VEP. The rare-variant *CXCR6* signal reached study-wide significance in the MANOLIS cohort (MAF=0.9%, β=1.51, σ=0.0215, $P=8.46x10^{-12}$) but was only observed as a singleton in the Pomak cohort. Interleukin 18 (IL18) is activated by NLRC4[51]. β: effect size, σ: standard error.

bioRxiv preprint doi: https://doi.org/10.1101/854752; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Protein	Indication	Disease definition	β	σ	Wald *P*
CHI3L1	Hypercholesterolemia	ICD code E780	0.604	0.135	7.57×10^{-6}
	Hypercholesterolemia	any E78 ICD code	0.639	0.130	9.77×10^{-7}
	High cholesterol	Self-reported	0.725	0.118	7.89×10^{-10}
	Hypercholesterolemia	Self-reported	0.604	0.135	7.57×10^{-6}
GRN	Hypercholesterolemia	ICD code E780	0.629	0.041	4.50×10^{-53}
	High cholesterol	Self-reported	0.762	0.036	1.69×10^{-99}
	Hypercholesterolemia	Self-reported	0.629	0.041	4.50×10^{-53}
PECAM1	Deep vein thrombosis	Self-reported	-2.47	0.170	7.39×10^{-48}
	High cholesterol	Self-reported	-0.71	0.078	6.90×10^{-20}
	Hypertension	Self-reported	0.410	0.060	1.26×10^{-11}
	Pulmonary embolism	ICD code I269	-2.81	0.230	2.80×10^{-34}
	Deep vein thrombosis	ICD code I802	-3.61	0.259	4.91×10^{-44}
	Pulmonary embolism	Self-reported	-3.22	0.254	1.02×10^{-36}

Table 2: Effect of protein level PRS in logistic disease prediction models in UK Biobank. Effects are reported for the most significantly associated protein scores only, statistics for the full disease-modelling regression including intercept and clinical covariates are given in Supplementary Table 11. Effect: log-odds change per score unit increase β : effect size, σ: standard error. Wald *P*: *P*-value for the Wald test of predictor significance. Threshold: *P* threshold used for constructing the score. SR: Self-reported. ICD: International Classification of Diseases, Tenth Revision.

bioRxiv preprint doi: https://doi.org/10.1101/854752; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figures

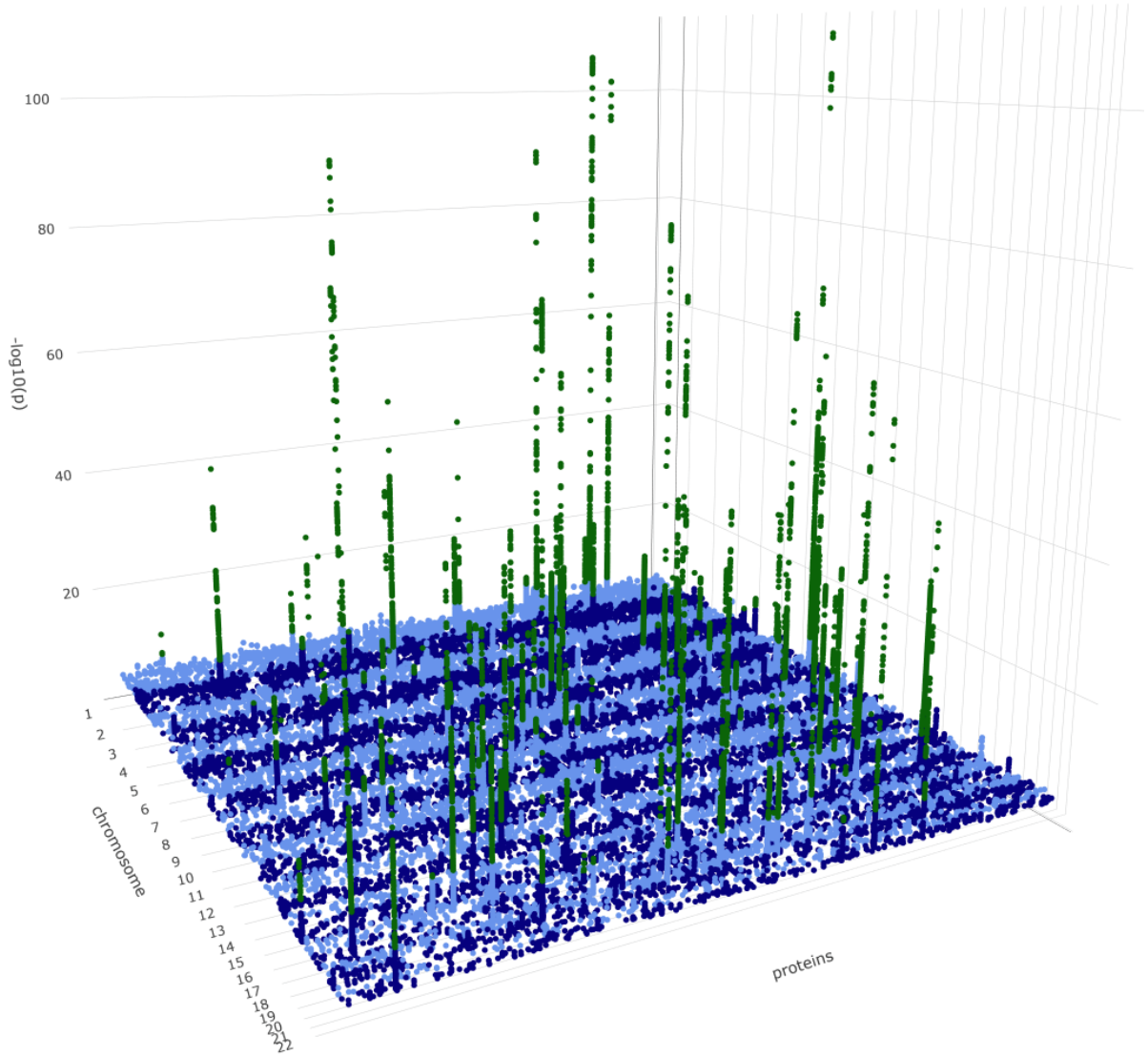


Figure 1 : Genome-wide association signals across all tested proteins. For clarity, variants with $P>1x10^{-5}$ are not represented in the figure. Variants with $P<7.45x10^{-11}$ are plotted in green.

bioRxiv preprint doi: https://doi.org/10.1101/854752; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

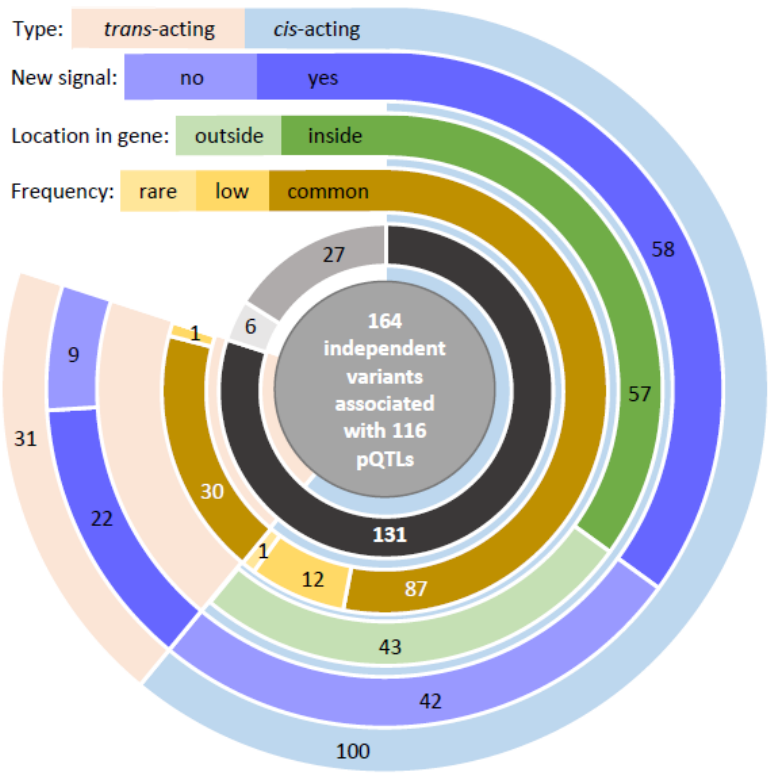


Figure 2: Characteristics of independently contributing pQTL variants. The innermost circle represents replication status: dark grey for variants that replicate, medium grey for variants that do not replicate and light grey for variants for which no proxy was found in the Pomak dataset.

bioRxiv preprint doi: https://doi.org/10.1101/854752; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

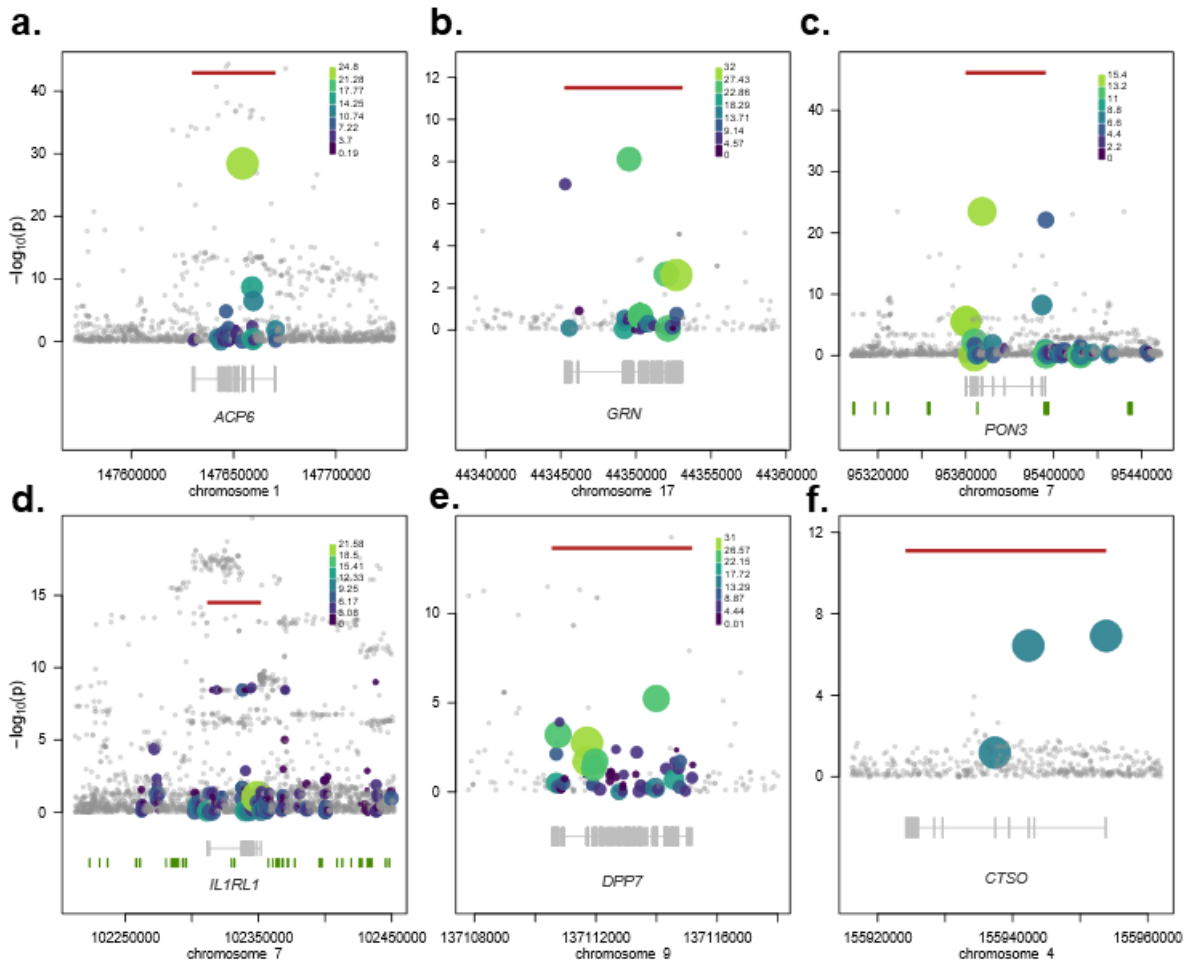
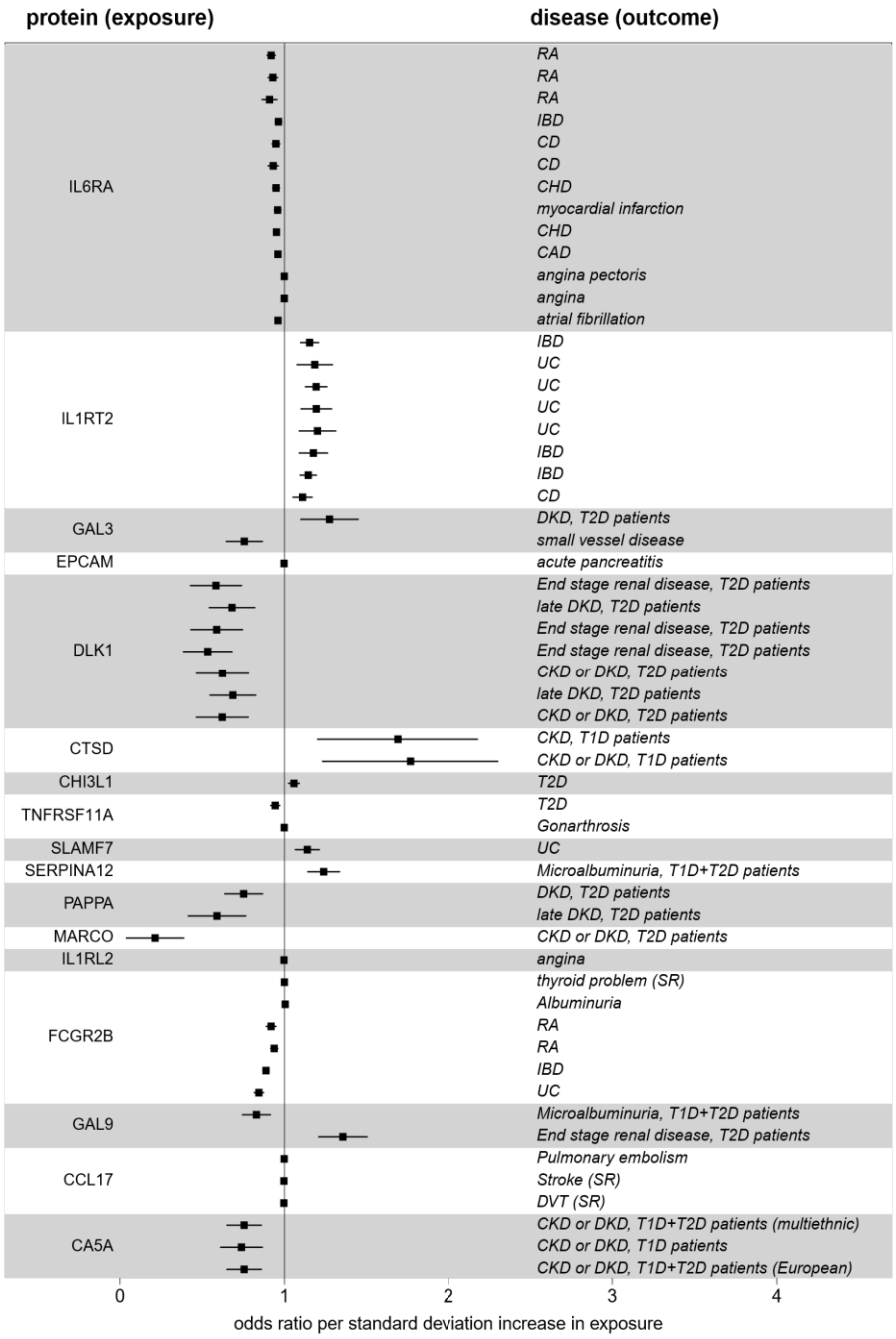


Figure 3: Rare variant pQTLs. Rare variant burden signals detected in this study -the most significant burden per gene is displayed. Circles denote the sequence variants identified in the region. Genes are denoted in gray below the regional association plots; bars represent exons across all transcripts. Horizontal red lines indicate the -log10 of the burden signal p-value, with size and colour of circles proportional to the weighting scheme used (CADD or Eigen). Grey circles denote variants not included in the burden. Details on variants included in each burden are given in Supplementary Table 12.


protein (exposure)
disease (outcome)
IL6RA
RA
RA
RA
IBD
CD
CD
CHD
myocardial infarction
CHD
CAD
angina pectoris
angina
atrial fibrillation
IL1RT2
IBD
UC
UC
UC
UC
IBD
IBD
CD
GAL3
DKD, T2D patients
small vessel disease
EPCAM
acute pancreatitis
DLK1
End stage renal disease, T2D patients
late DKD, T2D patients
End stage renal disease, T2D patients
End stage renal disease, T2D patients
CKD or DKD, T2D patients
late DKD, T2D patients
CKD or DKD, T2D patients
CTSD
CKD, T1D patients
CKD or DKD, T1D patients
CHI3L1
T2D
TNFRSF11A
T2D
Gonarthrosis
SLAMF7
UC
SERPINA12
Microalbuminuria, T1D+T2D patients
PAPPA
DKD, T2D patients
late DKD, T2D patients
MARCO
CKD or DKD, T2D patients
IL1RL2
angina
FCGR2B
thyroid problem (SR)
Albuminuria
RA
RA
IBD
UC
GAL9
Microalbuminuria, T1D+T2D patients
End stage renal disease, T2D patients
CCL17
Pulmonary embolism
Stroke (SR)
DVT (SR)
CA5A
CKD or DKD, T1D+T2D patients (multiethnic)
CKD or DKD, T1D patients
CKD or DKD, T1D+T2D patients (European)
0
1
2
3
4
odds ratio per standard deviation increase in exposure

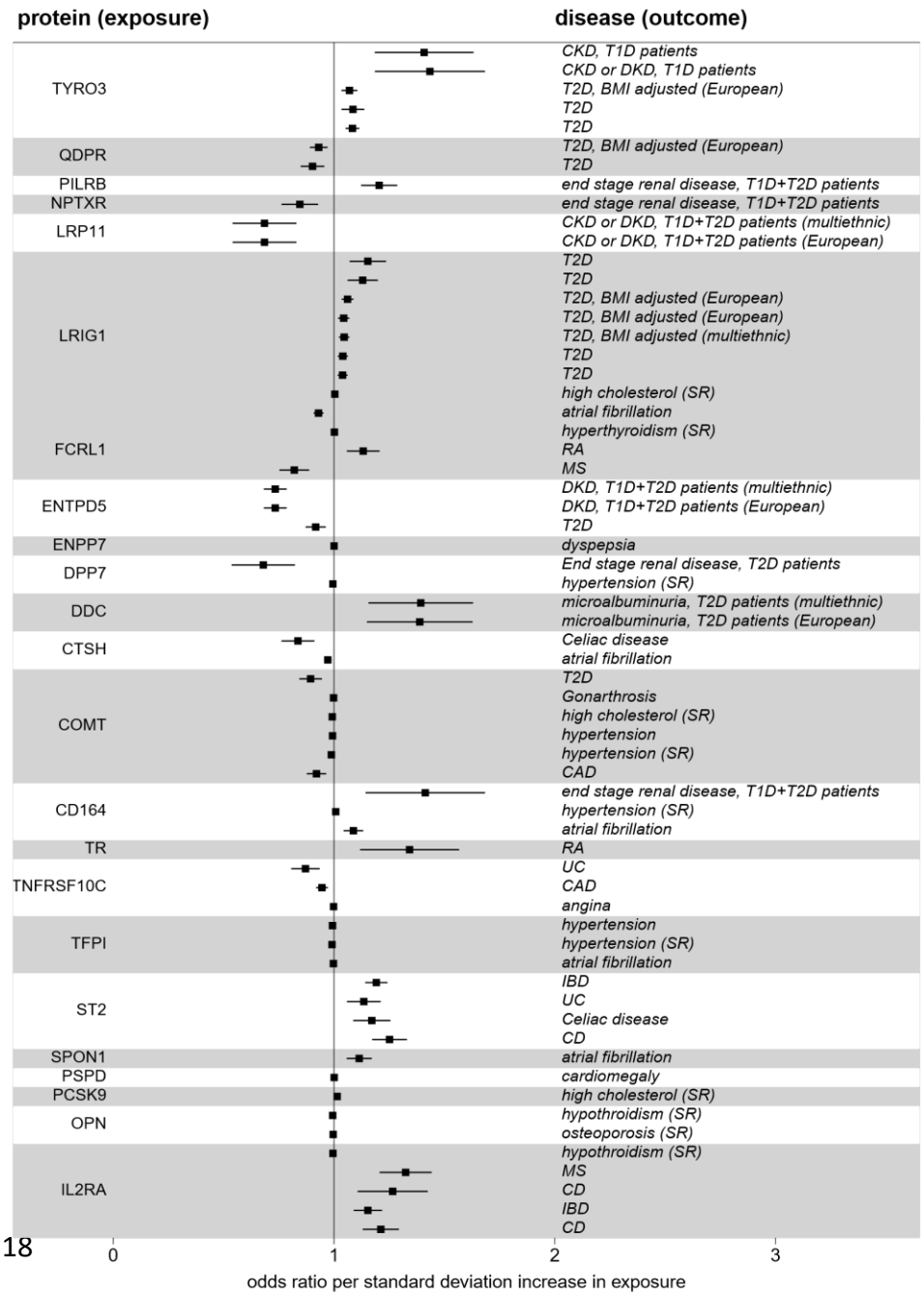

protein (exposure)
disease (outcome)
TYRO3
CKD, T1D patients
CKD or DKD, T1D patients
T2D, BMI adjusted (European)
T2D
T2D
QDPR
T2D, BMI adjusted (European)
T2D
PILRB
end stage renal disease, T1D+T2D patients
NPTXR
end stage renal disease, T1D+T2D patients
LRP11
CKD or DKD, T1D+T2D patients (multiethnic)
CKD or DKD, T1D+T2D patients (European)
LRIG1
T2D
T2D
T2D, BMI adjusted (European)
T2D, BMI adjusted (European)
T2D, BMI adjusted (multiethnic)
T2D
T2D
FCRL1
high cholesterol (SR)
atrial fibrillation
hyperthyroidism (SR)
RA
MS
ENTPD5
DKD, T1D+T2D patients (multiethnic)
DKD, T1D+T2D patients (European)
T2D
ENPP7
dyspepsia
DPP7
End stage renal disease, T2D patients
hypertension (SR)
DDC
microalbuminuria, T2D patients (multiethnic)
microalbuminuria, T2D patients (European)
CTSH
Celiac disease
atrial fibrillation
COMT
T2D
Gonarthrosis
high cholesterol (SR)
hypertension
hypertension (SR)
CAD
CD164
end stage renal disease, T1D+T2D patients
hypertension (SR)
atrial fibrillation
TR
RA
TNFRSF10C
UC
CAD
angina
TFPI
hypertension
hypertension (SR)
atrial fibrillation
ST2
IBD
UC
Celiac disease
CD
SPON1
atrial fibrillation
PSPD
cardiomegaly
PCSK9
high cholesterol (SR)
OPN
hypothroidism (SR)
osteoporosis (SR)
IL2RA
hypothroidism (SR)
MS
CD
IBD
CD
0
1
2
3
odds ratio per standard deviation increase in exposure

Figure 4: Significant causal protein-disease associations identified through two-sample Mendelian randomisation. Protein (exposure) names are indicated on the left, diseases (outcomes) on the right. Identical disease names for a given protein indicate a MR signal replicating across multiple studies of the same disease; further details and causal associations with quantitative traits are displayed in Supplementary Table 3. RA: rheumatoid arthritis, IBD: inflammatory bowel disease, CD: Crohn's disease, CHD: Coronary heart disease, CAD: Coronary artery disease, UC: ulcerative colitis, DKD: diabetic kidney disease, T2D: type 2 diabetes, T1D: type 1 diabetes, CKD: chronic kidney disease, MS: multiple sclerosis.

bioRxiv preprint doi: https://doi.org/10.1101/854752; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

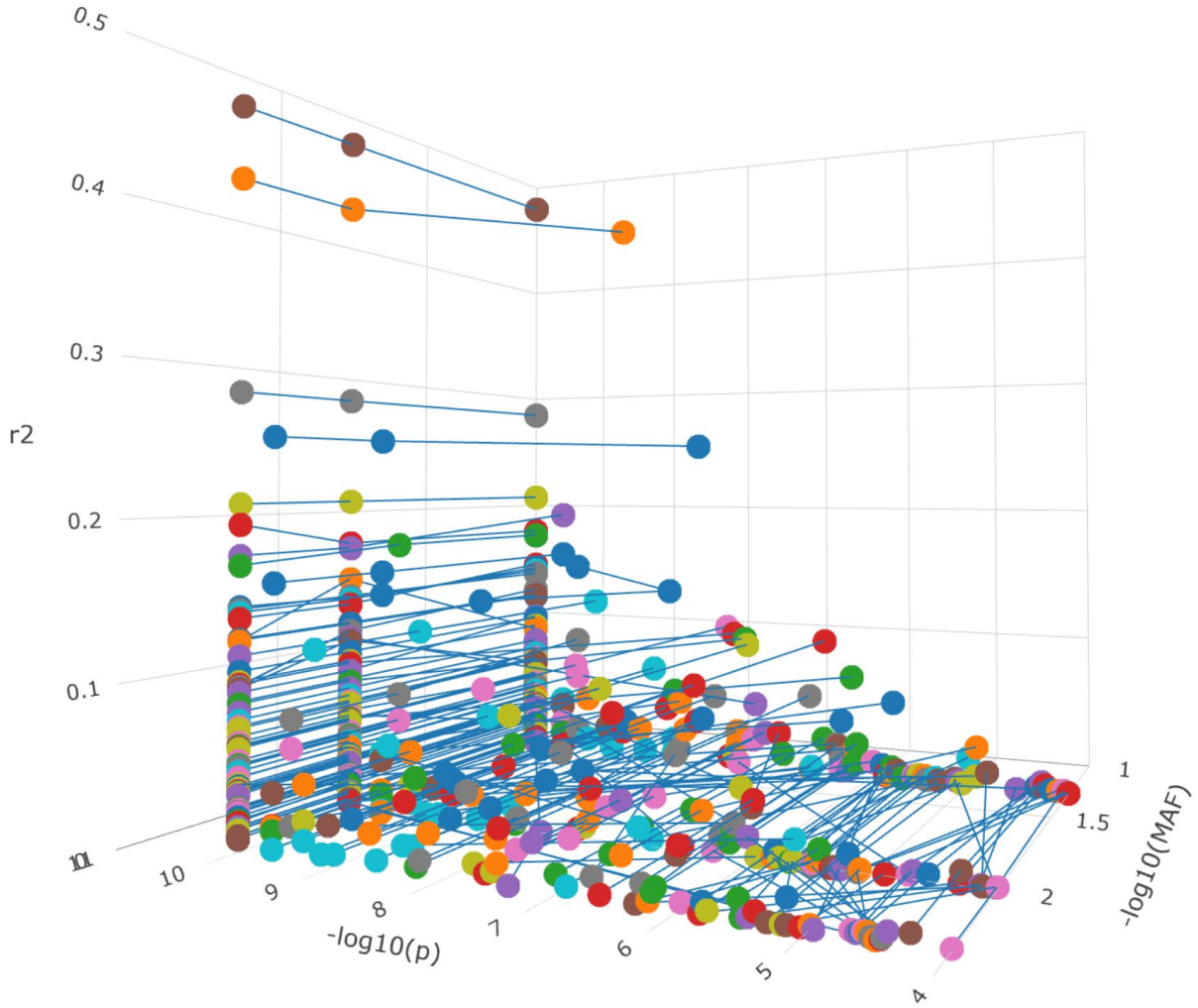


Figure 5: Variance explained by PRS for protein levels in the Pomak cohort. Scores calculated from the MANOLIS summary statistics, at $1x10^{-10}$ p-value threshold intervals, and for 3 different MAF cutoffs (5%, 1%, and 0.3%). Only the threshold providing the best r^2 is displayed, for each protein and MAF threshold. Values for the same protein are connected by blue lines. Most lines are contained within a z-plane, indicating that the same optimal r^2 can be achieved at all MAF thresholds. The 5 best performing scores are, from top to bottom: CHIT1, MEP1B, SHPS1, IL6RA, ICAM2 (Supplementary Table 13).

bioRxiv preprint doi: https://doi.org/10.1101/854752; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

1. Lundberg, M., Eriksson, A., Tran, B., Assarsson, E. & Fredriksson, S. Homogeneous antibody-based proximity extension assays provide sensitive and specific detection of low-abundant proteins in human blood. *Nucleic Acids Res* **39**, e102 (2011).
2. Southam, L. *et al.* Whole genome sequencing and imputation in isolated populations identify genetic associations with medically-relevant complex traits. *Nat Commun* **8**, 15606 (2017).
3. Suhre, K. *et al.* Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat Commun* **8**, 14357 (2017).
4. Folkersen, L. *et al.* Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. *PLoS Genet* **13**, e1006706 (2017).
5. Emilsson, V. *et al.* Co-regulatory networks of human serum proteins link genetics to disease. *Science* **361**, 769-773 (2018).
6. Sun, B.B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73-79 (2018).
7. Yao, C. *et al.* Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nat Commun* **9**, 3268 (2018).
8. Lu, Z.H. *et al.* The promiscuous chemokine binding profile of the Duffy antigen/receptor for chemokines is primarily localized to sequences in the amino-terminal domain. *J Biol Chem* **270**, 26239-45 (1995).
9. Kashiwazaki, M. *et al.* A high endothelial venule-expressing promiscuous chemokine receptor DARC can bind inflammatory, but not lymphoid, chemokines and is dispensable for lymphocyte homing under physiological conditions. *Int Immunol* **15**, 1219-27 (2003).
10. Shimaoka, T. *et al.* Cell surface-anchored SR-PSOX/CXC chemokine ligand 16 mediates firm adhesion of CXC chemokine receptor 6-expressing cells. *J Leukoc Biol* **75**, 267-74 (2004).
11. Rennert, P. *et al.* A soluble form of B cell maturation antigen, a receptor for the tumor necrosis factor family member APRIL, inhibits tumor cell growth. *J Exp Med* **192**, 1677-84 (2000).
12. Huai, Q. *et al.* Structure of human urokinase plasminogen activator in complex with its receptor. *Science* **311**, 656-9 (2006).
13. Johnson, L.A. *et al.* Apolipoprotein E-low density lipoprotein receptor interaction affects spatial memory retention and brain ApoE levels in an isoform-dependent manner. *Neurobiol Dis* **64**, 150-62 (2014).
14. Weiss, E.S. *et al.* Interleukin-18 diagnostically distinguishes and pathogenically promotes human and murine macrophage activation syndrome. *Blood* **131**, 1442-1455 (2018).
15. Maiolino, G. *et al.* The role of oxidized low-density lipoproteins in atherosclerosis: the myths and the facts. *Mediators Inflamm* **2013**, 714653 (2013).
16. Pei, C. *et al.* Emerging role of interleukin-33 in autoimmune diseases. *Immunology* **141**, 9-17 (2014).
17. Roselli, C. *et al.* Multi-ethnic genome-wide association study for atrial fibrillation. *Nat Genet* **50**, 1225-1233 (2018).
18. Nielsen, J.B. *et al.* Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nat Genet* **50**, 1234-1239 (2018).

bioRxiv preprint doi: https://doi.org/10.1101/854752; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

19. Bult, C.J. *et al.* Mouse Genome Database (MGD) 2019. *Nucleic Acids Res* **47**, D801-D806 (2019).
20. Overgaard, M.T. *et al.* Expression of recombinant human pregnancy-associated plasma protein-A and identification of the proform of eosinophil major basic protein as its physiological inhibitor. *J Biol Chem* **275**, 31128-33 (2000).
21. Oxvig, C. The role of PAPP-A in the IGF system: location, location, location. *J Cell Commun Signal* **9**, 177-87 (2015).
22. Bach, L.A. & Hale, L.J. Insulin-like growth factors and kidney disease. *Am J Kidney Dis* **65**, 327-36 (2015).
23. Read, R. *et al.* Ectonucleoside triphosphate diphosphohydrolase type 5 (Entpd5)-deficient mice develop progressive hepatopathy, hepatocellular tumors, and spermatogenic arrest. *Vet Pathol* **46**, 491-504 (2009).
24. Durst, M.A., Ratia, K. & Lavie, A. Identifying small molecule probes of ENTPD5 through high throughput screening. *PLoS One* **14**, e0210305 (2019).
25. Gustafsen, C. *et al.* The hypercholesterolemia-risk gene SORT1 facilitates PCSK9 secretion. *Cell Metab* **19**, 310-8 (2014).
26. Gilly, A. *et al.* Cohort-wide deep whole genome sequencing and the allelic architecture of complex traits. *Nat Commun* **9**, 4674 (2018).
27. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* **44**, 821-4 (2012).
28. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* **44**, 369-75, S1-3 (2012).
29. Lloyd-Jones, L.R. *et al.* The Genetic Architecture of Gene Expression in Peripheral Blood. *Am J Hum Genet* **100**, 228-237 (2017).
30. Jiang, D. & McPeek, M.S. Robust rare variant association testing for quantitative traits in samples with related individuals. *Genet Epidemiol* **38**, 10-20 (2014).
31. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-5 (2014).
32. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J.D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet* **48**, 214-20 (2016).
33. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).
34. Carithers, L.J. *et al.* A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreserv Biobank* **13**, 311-9 (2015).
35. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648-60 (2015).
36. Kamat, M.A. *et al.* PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. *Bioinformatics* (2019).
37. Koscielny, G. *et al.* Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res* **45**, D985-D994 (2017).
38. Wishart, D.S. *et al.* DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* **36**, D901-6 (2008).
39. Sandholm, N. *et al.* The Genetic Landscape of Renal Complications in Type 1 Diabetes. *J Am Soc Nephrol* **28**, 557-574 (2017).
40. Willer, C.J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat Genet* **45**, 1274-1283 (2013).

bioRxiv preprint doi: https://doi.org/10.1101/854752; this version posted November 29, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

41. Morris, A.P. *et al.* Trans-ethnic kidney function association study reveals putative causal genes and effects on kidney-specific disease aetiologies. *Nat Commun* **10**, 29 (2019).
42. Haas, M.E. *et al.* Genetic Association of Albuminuria with Cardiometabolic Disease and Blood Pressure. *Am J Hum Genet* **103**, 461-473 (2018).
43. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in approximately 700000 individuals of European ancestry. *Hum Mol Genet* **27**, 3641-3649 (2018).
44. Moskvina, V. & Schmidt, K.M. On multiple-testing correction in genome-wide association studies. *Genet Epidemiol* **32**, 567-73 (2008).
45. Cheverud, J.M. A simple correction for multiple comparisons in interval mapping genome scans. *Heredity (Edinb)* **87**, 52-8 (2001).
46. Choi, S.W. & O'Reilly, P.F. PRSice-2: Polygenic Risk Score software for biobank-scale data. *Gigascience* **8**(2019).
47. Ahola-Olli, A.V. *et al.* Genome-wide Association Study Identifies 27 Loci Influencing Concentrations of Circulating Cytokines and Growth Factors. *Am J Hum Genet* **100**, 40-50 (2017).
48. Matteini, A.M. *et al.* Novel gene variants predict serum levels of the cytokines IL-18 and IL-1ra in older adults. *Cytokine* **65**, 10-6 (2014).
49. Higgins, J.P. & Thompson, S.G. Quantifying heterogeneity in a meta-analysis. *Stat Med* **21**, 1539-58 (2002).
50. Higgins, J.P., Thompson, S.G., Deeks, J.J. & Altman, D.G. Measuring inconsistency in meta-analyses. *BMJ* **327**, 557-60 (2003).
51. Canna, S.W. *et al.* Life-threatening NLRC4-associated hyperinflammation successfully treated with IL-18 inhibition. *J Allergy Clin Immunol* **139**, 1698-1701 (2017).