Connecting genetic risk to disease end points through the human blood plasma proteome

Karsten Suhre1**, Matthias Arnold2,*, Aditya Mukund Bhagwat3,*, Richard J. Cotton3,*, Rudolf Engelke3,*, Johannes Raffler2,*, Hina Sarwath3,*, Gaurav Thareja1,*, Annika Wahl4,5,*, Robert Kirk DeLisle6, Larry Gold6, Marija Pezer7, Gordan Lauc7, Mohammed A. El-Din Selim8, Dennis O. Mook-Kanamori9, Eman K. Al-Dous10, Yasmin A. Mohamoud10, Joel Malek10, Konstantin Strauch11,12, Harald Grallert4,5,13, Annette Peters5,13,14, Gabi Kastenmüller2,13, Christian Gieger4,5,13,** & Johannes Graumann3,***

Genome-wide association studies (GWAS) with intermediate phenotypes, like changes in metabolite and protein levels, provide functional evidence to map disease associations and translate them into clinical applications. However, although hundreds of genetic variants have been associated with complex disorders, the underlying molecular pathways often remain elusive. Associations with intermediate traits are key in establishing functional links between GWAS-identified risk-variants and disease end points. Here we describe a GWAS using a highly multiplexed aptamer-based affinity proteomics platform. We quantify 539 associations between protein levels and gene variants (pQTLs) in a German cohort and replicate over half of them in an Arab and Asian cohort. Fifty-five of the replicated pQTLs are located in trans. Our associations overlap with 57 genetic risk loci for 42 unique disease end points. We integrate this information into a genome-proteome network and provide an interactive web-tool for interrogations. Our results provide a basis for novel approaches to pharmaceutical and diagnostic applications.

1 Department of Physiology and Biophysics, Weill Cornell Medicine-Qatar, Education City, PO 24144 Doha, Qatar. 2 Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany. 3 Proteomics Core, Weill Cornell Medicine-Qatar, Education City, PO 24144 Doha, Qatar. 4 Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany. 5 Institute of Epidemiology II, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany. 6 SomaLogic, 2945 Wilderness Pl, Boulder, Colorado 80301, USA. 7 Genos Ltd, Glycoscience Research Laboratory, Hrđkovca 2/11, 10000 Zagreb, Croatia. 8 Department of Dermatology, Hamad Medical Corporation, PO Box 3050 Doha, Qatar. 9 Leiden University Medical Centre, PO Box 9600, 2300 RC Leiden, The Netherlands. 10 Genomics Core, Weill Cornell Medicine-Qatar, Education City, PO 24144, Doha, Qatar. 11 Institute of Genetic Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany. 12 Institute of Medical Informatics, Biometry and Epidemiology, Chair of Genetic Epidemiology, Ludwig-Maximilians-Universität, Marchioninistrasse 15, 81377 München, Germany. 13 German Center for Diabetes Research (DZD), Ingolstädter Landstraße 1, 85764 Neuherberg, Germany. 14 German Center for Cardiovascular Disease Research (DZHK), Oudenander Straße 16, 13347 Berlin, Germany.

* These authors contributed equally to this work. ** These authors jointly supervised this work. † Present address: Scientific Service Group Biomolecular Mass Spectrometry, Max Planck Institute for Heart and Lung Research, W.G. Kerckhoff Institute, Ludwigstr. 43, D-61231 Bad Nauheim, Germany. Correspondence and requests for materials should be addressed to K.S. (email: karsten@suhre.fr) or to C.G. (email: christian.gieger@helmholtz-muenchen.de) or to J.G. (email: johannes.graumann@mpi-bn.mpg.de).
Genome-wide association studies (GWAS) with intermediate phenotypes, such as metabolite and protein levels, can reveal variations in protein abundances, enzymatic activities, interaction properties or regulatory mechanisms, which can inform clinical applications1–3. Studies with aptamer-, immunoassay- and mass-spectrometry-based proteomics (Supplementary Table 1) have shown that protein levels are strongly modulated by variations in the nearby cis regions of their encoding genes4–9. However, proteomics-based genetic association studies have been limited by small protein panels or cohort sizes, and thus, have not systematically addressed associations across multiple chromosome locations. These trans-associations are key in establishing functional links between different risk-variants, because they reveal the downstream effectors in the pathways between cis-encoded risk-variants and disease end points.

Here we use a highly multiplexed, aptamer-based, affinity proteomics platform (SOMAscan)10 in a GWAS to quantify levels of 1,124 proteins in blood plasma samples. We investigate samples from 1,100 individuals of the population-based KORA (Cooperative Health Research in the Region of Augsburg) study1,11 and replicate the results in 338 participants of the Qatar Metabolomics Study on Diabetes (QMDiab) with participants of Arab and Asian ethnicities12.

We identify 539 independent, single-nucleotide polymorphism (SNP)-protein associations, which include novel intrachromosomal (trans) links related to autoimmune disorders, Alzheimer’s disease, cardiovascular disease, cancer and many other disease end points. Our study represents the outcome of over 1,100 new GWAS with blood circulating protein levels, many of which have never been reported before. We find that up to 60% of the naturally occurring variance in the blood plasma levels of essential proteins can be explained by two or more independent variants of a single gene located on another chromosome, and identify strong genetic associations with intermediate traits in pathways involved in complex disorders.

**Results**

**A GWAS with blood circulating protein levels.** We used linear additive genetic regression models, adjusted for relevant covariates, to analyse 509,946 common autosomal SNPs for genome-wide associations with 1,124 protein levels measured in 1,000 blood samples from the KORA study (see Methods). We grouped association signals with correlated SNPs ($r^2 > 0.1$) into single-protein quantitative trait loci (pQTLs) and identified 539 pQTLs (284 unique proteins, 451 independent SNPs) with a conservative Bonferroni level of significance ($P < 8.72 \times 10^{-11}$ = 0.05/509,946/1,124; unless otherwise stated we report uncorrected $P$ values for the reported traits from linear regression models). For 21% of the assayed proteins we detected a cis-pQTL at a more liberal significance cutoff of $P < 5 \times 10^{-8}$. The full list is available as Supplementary Data 1. We then fine-mapped our associations with variants imputed from the 1,000-Genomes project database. Next, we attempted replication of 462 associations that had a suitable genotype tag SNP ($r^2 > 0.8$) in samples from 338 participants of the QMDiab study12. Of those, 234 associations (50.6%) were replicated at a Bonferroni level of significance ($P < 1.08 \times 10^{-4}$ = 0.05/462), and an additional 150 (totalling to 83.1%) showed nominal significance ($P < 0.05$) in the replication sample. We observed directional consistency between primary and replication sample for 215 of the 234 replicated SNPs. Discrepancy for the remaining 19 SNPs may be explained by changes in major and minor allele coding between the two cohorts and ethnic differences. What is more, out of 208 associations that had 95% replication power (determined by sampling), 171 (82.2%) were replicated and 198 (95.2%) were nominally significant. Moreover, we replicated several associations previously reported in aptamer-, immunoassay- and mass-spectrometry-based studies, which demonstrated concordance among these technologies (Supplementary Tables 1–5).

**Interconnectedness and annotation of the plasma proteome.** The SOMAscan aptamers were generated to bind specifically to proteins implicated in numerous diseases and physiological processes and to target a broad range of secreted, intracellular and extracellular proteins, which are detectable in blood plasma (Supplementary Fig. 1, Supplementary Data 2). To identify interrelations between these proteins, we used Gaussian graphical modelling (GGM) to connect 1,092 proteins through 3,943 Bonferroni-significant partial protein–protein correlation edges (Supplementary Data 3). Then, we added all 451 genetic pQTL variants as nodes, and connected them to the protein network through the 539 SNP-protein associations (Fig. 1). This network is freely available online, and it can be navigated via an interactive web interface. Overall, we found that given our study power the blood plasma levels of over 20% of all assayed proteins were under substantial genetic control, and in some cases, this control resulted in nearly total protein ablation (Fig. 2).

Additionally, a wide-spread feature was the convergence of multiple association signals to impact the levels of key proteins of biomedical and pharmaceutical interest (Supplementary Fig. 2): with the Ingenuity Pathway Analysis database (IPA, Qiagen Inc.), we annotated 50 proteins as clinical or pharmaceutical biomarkers (Supplementary Table 6) and 43 as drug targets (Supplementary Table 7), and all of these had at least one replicated pQTL.

Next, we identified and annotated putative causative and disease-relevant variants. We used the SNiPA web-tool13 to retrieve variant-specific annotations for all SNPs from the 1,000-Genomes project that were in linkage disequilibrium with an identified pQTL (linkage disequilibrium $r^2 > 0.8$). The annotations included primary effect predictions, SNPs in experimentally identified regulatory elements (ENCODE), expression QTLs (eQTLs) and disease associated variants. Of 384 annotated cis-pQTLs, 228 had a variant in the gene coding region, whereof 74 were protein-changing. Eighty-eight had a variant in a regulatory element, and 179 had an eQTL that matched the associated protein. We complemented the pQTL annotation with 122 overlapping methylation QTLs (meQTLs) (Supplementary Data 1) and 14 overlapping metabolic QTLs (mQTLs) (Supplementary Table 8). With the GWAS catalogue as a reference, and by including publicly available summary statistics data from 16 large disease GWAS consortia, we identified 83 GWAS associations for 42 unique disease end points that overlapped with 57 pQTL loci (Supplementary Table 9).

**Trans-pQTLs link genotype to GWAS end points.** Trans-associations are exceptionally valuable for identifying new pathways. These associations establish causal links between proteins encoded at the GWAS loci and the blood levels of one or several trans-encoded proteins. We identified 148 trans-pQTLs and replicated 55. Forty-nine trans-pQTLs had 95% replication power, we replicated 38 of these. Six replicated trans-pQTLs had an additional replicated cis-association, two had two replicated trans-associations and one had three replicated trans-associations. Three independent SNPs at the haptoglobin (HP) locus had together four pQTLs, and
two proteins had replicated trans-associations at two distinct chromosome locations (Tables 1 and 2).

In this study, the pleiotropic ABO blood group gene exhibited the most promiscuous protein association signal. This locus displayed six independent genetic variants through replicated trans-pQTLs; three of these associations had been published previously (VWF, SELE, SELP, Supplementary Table 2). The other eleven associations were new, to our knowledge (BCAM, CD200, CD209, CDH5, SELP, Supplementary Table 2). The other eleven associations had been published previously (VWF, SELE, SELP, Supplementary Table 2).

The non-O blood group is one of the most important genetic risk factors for venous thromboembolism,16 pancreatic cancer,17 and susceptibility to infectious diseases.18 However, we lack a full understanding of the proteins and pathways involved in the pathogenic effects of these genetic variants. Based on the pQTLs reported here, we found support for the following three new hypotheses: First, the association between ABO and levels of circulating insulin receptor (INSR) reflects the previously reported associations between the ABO locus and diabetes and the insulin receptor and diabetes.15 The association between ABO and the insulin receptor suggests that INSR-mediated insulin signalling may be involved in the ABO-diabetes

Figure 1 | The genome-proteome-disease network. (a) Data sources integrated into the network, indicating the number and type of the overlapping associations, from the SNP to the disease end point; all associations are freely accessible at http://proteomics.gwas.eu. (b) Circular plot of all cis- and trans-associations, cis-pQTLs are indicated by triangles, trans-pQTLs connect associated variant locations and trans-encoded protein locations. An interactive version of this circular plot constitutes an entry point to query the integrated web-server. (c) Example of a genome-proteome-disease sub-network obtained from the server for a query using the search word ‘Crohn’s Disease’. Network elements are disease traits (pink hexagons), pQTL loci (green diamonds), protein levels (blue ovals); nodes are connected by genetic associations, partial correlations and disease GWAS associations. This example (edited here for clarity) revealed four risk loci that associated with plasma levels of C7, MST, IL23R and IL18R, respectively. These four proteins all have a major role in auto-immune disorders. Partial correlations between neighbouring proteins reveal pathways that may be involved in the aetiology of Crohn’s disease. Similar networks can be retrieved starting with a query using any of the 539 pQTLs, 1,124 proteins and 42 unique co-associated disease endpoints. In the integrated web-server, all items are interactively linked to association data from the discovery and the replication study, regional association plots based on imputed variants, locus annotations including co-associated eQTL-, meQTL-, mQTL-, regulatory-, coding- and disease risk-variants, and link-outs to relevant protein databases, original data sources and primary publications. The links in this network reflect the outcome of many natural experiments, represented by genetic variations observed in the genomes of hundreds of individuals from the general population and probed by deep proteomics phenotyping using over 1,100 different aptamers.
association. Second, rs651007 associated here with P-selectin (SELP). SELP-positive platelets were previously reported to be associated with blood pressure\textsuperscript{19}, and rs651007 was associated in a GWAS with angiotensin converting enzyme activity. This variant was further identified as an mQTL for a number of dipeptides in blood\textsuperscript{20}, which may be produced by the dipeptidase angiotensin converting enzyme\textsuperscript{21}. These observations suggest a potential role for the SELP pQTL in the GWAS association between ABO and cardiovascular disease. Third, several of the 14 proteins associated here with ABO have been shown to interact or form complexes in relation with angiogenesis and vascular maturation processes: The angiopoietin-1 receptor (TEK) has or form complexes in relation with angiogenesis and vascular metabolism\textsuperscript{22,23}.

**Post-translational modifications.** To follow up on the many hypotheses generated by the pQTLs reported here, expert knowledge and experimentation will be required. Given the fact that several of the proteins identified in our pQTLs were glycoproteins, we performed plasma protein glycoprofiling (see Online Methods) to investigate whether some of our pQTLs were involved in post-translational protein modifications.

For instance, SNP rs3760775, located at the FUT3 gene locus, was associated with plasma levels of the there encoded galactose-\(\alpha\)-side3(4)-L-fucosyltransferase protein. We found that this same variant was strongly associated with the \(N\)-glycan GP33 (\(P = 1.4\) \(\times 10^{-15}\)) (see Fig. 3a for the GP33 glycan structure). In a previous GWAS, SNP rs3760775 was reported to be associated with the glycan antigen, CA19-9 (ref. 29), a widely used cancer biomarker, which is present on multiple proteins\textsuperscript{30}. Nearly 20 years ago, studies demonstrated a similar association between CA19-9 and the FUT3 gene dosage\textsuperscript{31,32}. This previously reported association between variance in the FUT3 locus and the expression of cancer antigen, CA19-9, and our observation of a similar association with GP33, suggest that both glycans may be involved in a same pathway.

Another glycoprofile investigation began with the strong, genotype-dependent correlation between complement factor C4 and the \(N\)-glycan, GP19 (\(P = 2.0\) \(\times 10^{-27}\)) (Fig. 3d). We found that SNP rs82823 was associated with both C4 and GP19. GP19 is composed of nine mannose moieties (M9 glycan structure), and it was previously reported to be attached to GP33 (\(P = 1.4\) \(\times 10^{-15}\)) (see Fig. 3a for the GP33 glycan structure). In a previous GWAS, SNP rs3760775 was reported to be associated with the glycan antigen, CA19-9 (ref. 29), a widely used cancer biomarker, which is present on multiple proteins\textsuperscript{30}. Nearly 20 years ago, studies demonstrated a similar association between CA19-9 and the FUT3 gene dosage\textsuperscript{31,32}. This previously reported association between variance in the FUT3 locus and the expression of cancer antigen, CA19-9, and our observation of a similar association with GP33, suggest that both glycans may be involved in a same pathway.

**Biomedical relevance.** A major challenge in conducting a disease GWAS is the difficulty in identifying causative variants in the pathophysiological pathways that lead to the observed clinical manifestations. Generally, hypotheses about the identity of the disease-causing genes are based on biological arguments, such as the presence of SNPs in the regulatory or coding regions of functionally plausible genes, which may be supported by co-associated eQTLs. However, a stronger argument is provided by the co-association with a pQTL, which constitutes firm experimental evidence that the blood levels of
| Locus | Candidate cis-gene(s) | Sentinel SNP | Chr | Position | Type | P value KORA | P value QMDiab | Protein name (gene symbol) |
|-------|----------------------|--------------|-----|----------|------|--------------|---------------|--------------------------|
| 192   | GJA9<sup>ace</sup>, RHBDL2<sup>ace</sup>, RPS-864K19.6<sup>ace</sup>, RRAGC<sup>ab</sup>, MYCBp<sup>ace</sup>, ... | rs4494141 1 | 39,339,682 | Trans | 1.9 × 10^-20 | 3.0 × 10^-8 | Kunitz-type protease inhibitor 1 (SPINT1) |
| 210   | Clec16B<sup>bc</sup>, C8A<sup>a</sup>, C8B<sup>a</sup>, DAB1<sup>b</sup> | rs626457 1 | 57,407,484 | Trans | 3.3 × 10^-19 | 1.7 × 10^-8 | Neurexophilin-1 (NXPH1) |
| 71    | PSRC1<sup>bc</sup>, CELSR2<sup>ac</sup>, AMPD2<sup>bc</sup>, SORTC1<sup>c</sup> | rs646776 1 | 109,818,330 | Trans | 1.0 × 10^-52 | 1.7 × 10^-18 | Granulins (GRN) |
| 413   | F5<sup>ae</sup> | rs9332653 1 | 169,490,772 | Trans | 1.6 × 10^-11 | 1.5 × 10^-5 | Coagulation Factor V (F5) |
| 17    | F5<sup>ae</sup>, SELP<sup>b</sup>, SELL<sup>b</sup> | rs4525 1 | 169,511,734 | Trans | 2.0 × 10^-11 | 1.3 × 10^-32 | Calcium/calmodulin-dependent protein kinase type 1 (CAMK1) |
| 98    | CH3, CHFR<sup>c</sup> | rs6695321 1 | 196,675,861 | Trans | 9.4 × 10^-40 | 5.1 × 10^-5 | Calcium/calmodulin-dependent protein kinase type 1 (CAMK1) |
| 72    | CFHR4<sup>ae</sup>, CFHR2<sup>a</sup>, ASPM<sup>a</sup>, ZBTB14<sup>b</sup> | rs10494745 1 | 196,887,457 | Trans | 1.8 × 10^-52 | 2.4 × 10^-11 | Complement C1s subcomponent (C1S) |
| 76    | CFHR4<sup>ae</sup>, CFHR2/5<sup>a</sup>, CFHR1/3<sup>c</sup>, CH3 | rs10801582 1 | 196,944,357 | Trans | 1.1 × 10^-49 | 2.0 × 10^-11 | Haemopexin (HPX) |
| 122   | TRIM5<sup>ae</sup> | rs1339847 1 | 248,039,294 | Trans | 9.2 × 10^-33 | 5.4 × 10^-6 | Haemopexin (HPX) |
| 93    | COLEC11<sup>ae</sup>, ALLC<sup>c</sup> | rs7588285 2 | 3,648,186 | Trans | 1.1 × 10^-40 | 1.4 × 10^-16 | Dynin light chain block-type 1 (DYNLRB1) |
| 157   | LTF<sup>abe</sup>, CCR5<sup>b</sup>, CCR3<sup>bc</sup> | rs1126478 3 | 46,501,213 | Trans | 2.7 × 10^-37 | 8.7 × 10^-20 | Interleukin-19 (IL19) |
| 95    | IPK2<sup>b</sup>, CELSR3<sup>bc</sup>, NCKIPSD<sup>bc</sup>, ARH5<sup>abc</sup>, USP9<sup>ab</sup> | rs1171583 3 | 48,770,732 | Trans | 2.2 × 10^-40 | 4.2 × 10^-8 | Alkaline phosphatase, tissue-nonspecific isozyme (ALPL) |
| 91    | RRM6<sup>ac</sup>, RNF123<sup>bc</sup>, B5N<sup>a</sup>, AMIGO2<sup>c</sup>, GMPPB<sup>b</sup>, ... | rs4688759 3 | 50,008,118 | Trans | 1.8 × 10^-42 | 4.2 × 10^-8 | Thioredoxin domain-containing protein 12 (TXNDC12) |
| 52    | DCBLD2<sup>c</sup>, CPOX<sup>c</sup>, ST3GAL6<sup>c</sup> | rs10953480 3 | 98,431,986 | Trans | 9.9 × 10^-70 | 1.2 × 10^-17 | Thioredoxin domain-containing protein 12 (TXNDC12) |
| 352   | DNAJC13<sup>abc</sup>, ACAD11<sup>ab</sup>, NPH3<sup>bc</sup>, ACR<sup>a</sup>, UBA5<sup>a</sup>, ... | rs17412738 3 | 132,257,419 | Trans | 5.3 × 10^-13 | 3.8 × 10^-5 | Vascular endothelial growth factor receptor 3 (FLT4) |
| 203   | PCOCLE2<sup>ac</sup>, U2SURP<sup>b</sup>, ATR<sup>d</sup>, PL5<sup>d</sup> | rs4683702 3 | 142,617,138 | Trans | 2.0 × 10^-19 | 8.3 × 10^-5 | C-C motif chemokine 21 (CCL21) |
| 31    | HRG<sup>ae</sup> | rs2228243 3 | 186,395,113 | Trans | 4.7 × 10^-24 | 6.0 × 10^-18 | Endothelin-converting enzyme 1 (ECE1) |
| 43    | HRG<sup>ae</sup> | rs1042445 3 | 186,395,436 | Trans | 8.4 × 10^-78 | 1.2 × 10^-22 | Histidine-rich glycoprotein (HRG) |
| 26    | KNG1<sup>ae</sup> | rs2304456 3 | 186,445,052 | Trans | 2.9 × 10^-97 | 2.8 × 10^-40 | Dual specificity mitogen-activated protein kinase 4 (MAP2K4) |
| 96    | KNG1<sup>ae</sup> | rs5030062 3 | 186,454,180 | Trans | 3.0 × 10^-40 | 1.7 × 10^-13 | Dual specificity mitogen-activated protein kinase 4 (MAP2K4) |
| 123   | SKIV2L<sup>ae</sup>, C2<sup>a</sup>, NELFE<sup>a</sup>, DOX<sup>a</sup>, STK19<sup>a</sup> | rs9283893 6 | 31,897,219 | Trans | 1.8 × 10^-35 | 8.0 × 10^-10 | Kininogen-1 (KNG1) |
| 159   | SKIV2L<sup>ace</sup>, C4B<sup>be</sup>, TNXB<sup>ab</sup>, DOX<sup>a</sup>, STK19<sup>a</sup>, ... | rs387608 6 | 31,914,557 | Trans | 2.5 × 10^-25 | 2.4 × 10^-8 | Leucine carboxyl methyltransferase 1 (LCMT1) |
| 182   | TAP2<sup>c</sup>, PSMB9<sup>a</sup>, TAP1<sup>a</sup>, PSMB8<sup>a</sup>, COL11A2<sup>b</sup>, ... | rs17220241 6 | 32,822,244 | Trans | 9.6 × 10^-22 | 1.3 × 10^-5 | Coagulation Factor XI (F11) |
| 417   | ZFP1M2<sup>d</sup>, CXCL5<sup>d</sup> | rs16873418 8 | 106,592,145 | Trans | 1.9 × 10^-11 | 1.5 × 10^-5 | Coagulation Factor XI (F11) |
| 442   | ABO<sup>f</sup>, SURF1<sup>f</sup>, SLCA26<sup>f</sup>, GGTT1<sup>c</sup> | rs7857390 9 | 136,128,546 | Trans | 5.8 × 10^-11 | 1.0 × 10^-6 | Plasma kallikrein (KLKB1) |
| 69    | ABO<sup>bce</sup>, OBP2<sup>bc</sup>, DBH<sup>b</sup>, SURF1/2<sup>b</sup>, ADAMTS2<sup>bc</sup>, ... | rs8176749 9 | 136,131,188 | Trans | 6.1 × 10^-53 | 2.7 × 10^-42 | Neutrophil collagenase (MMP8) |
| 354   | ABO<sup>sc</sup>, SURF<sup>f</sup> | rs8176720 9 | 136,132,873 | Trans | 7.4 × 10^-10 | 6.8 × 10^-8 | gp41 C34 peptide, HIV (Human-virus) |

Please see the legend of Table 2 for a description of the replicated trans-pQTL loci.

*NOTCH1 is encoded in cis on chromosome 9, but distant from ABO.
cause of inflammatory arthritis and affects one in 200 Europeans. Auto-immune disorders identified in previous GWAS reveal new insights into multiple disease-associated pathways following sections, we show how pQTLs identified in this study hypotheses generated from pQTLs (see example in Fig. 1c). In the genome. Moreover, partial correlations to functionally related protein levels and estimated replication power. Candidate genes for the protein associations were annotated by considering the -encoded target protein(s). eQTL (superscript c), a variant is a variant hits a regulatory element of that gene (superscript b), a trans-eQTL (superscript d), a variant is protein changing (superscript e). The list of candidate genes in this table is limited to the five most plausible candidate genes for each locus. The full list is available online and in Supplementary Data 1. Every Loci that comprise at least one replicated association. Loci are referenced in this study by numbers ranging from 1 to 451 (strongest to weakest) and sorted here by chromosome position.

| Locus | Candidate cis-gene(s) | Sentinel SNP | Chr | Position | Type | Pr value KORA | P value QMDiab | Protein name (gene symbol) |
|-------|-----------------------|-------------|-----|----------|------|--------------|-----------------|--------------------------|
| 36    | ABOαβε, TSC1γ, AK8β, SARDHβ, GBGT1ε | rs505922 | 9   | 136,149,229 | Trans | 1.2 × 10⁻²⁰ | 1.0 × 10⁻⁸ | von Willebrand factor (VWF) |
| 269   | ABOαε, DBHβ, ADAMTS2λ, SARDHγ, RALGDSζ,... | rs630510 | 9   | 136,149,581 | Trans | 7.6 × 10⁻⁸⁶ | 4.1 × 10⁻²⁸ | CD209 antigen (CD209) |
| 28    | ABOαε, DBHβ, ADAMTS2λ, SARDHγ, RALGDSζ,... | rs651007 | 9   | 136,153,875 | Trans | 1.6 × 10⁻¹⁵ | 2.4 × 10⁻¹⁰ | Tyrosine-protein kinase receptor Tie-1, soluble (TIE1) |
| 28    | ABOαε, DBHβ, ADAMTS2λ, SARDHγ, RALGDSζ,... | rs651007 | 9   | 136,153,875 | Trans | 1.2 × 10⁻⁹⁶ | 8.2 × 10⁻²³ | E-Selectin (SELE) |
| 79    | CENPFβ, HIF1ANε | rs7091871 | 10  | 101,810,304 | Trans | 1.0 × 10⁻⁴⁸ | 3.5 × 10⁻¹⁶ | Calcium/calmodulin-dependent protein kinase type 1 (CAMK1) |
| 150   | SK3αβ, S1DT2αβ, PCSK7β, BUD13β, RNF214β,... | rs12099358 | 11  | 116,726,048 | Trans | 1.4 × 10⁻¹² | 1.7 × 10⁻⁷ | Calpastatin (CAST) |
| 65    | OATαβββ, POU2F3αε, ARHGFT12β, TMEM136β, TRIM29β,... | rs692804 | 11  | 120,099,368 | Trans | 1.1 × 10⁻⁵⁶ | 5.3 × 10⁻²⁴ | Interleukin-25 (ILZ25) |
| 115   | C15αβεβ, C1Rβ | rs21476727 | 12  | 7,170,336 | Cis | 3.4 × 10⁻³⁵ | 3.6 × 10⁻⁷ | Complement C1r subcomponent (C1R) |
| 304   | POCI1β, GALNT4α, POCI1β, ATPB1ε | rs722414 | 12  | 89,937,437 | Trans | 2.3 × 10⁻¹⁴ | 3.1 × 10⁻⁶ | Complement C1q subcomponent (C1QA, C1QB, C1QC) |
| 8     | ZC3H13βε, CP28ε, LCP1δ | rs1926447 | 13  | 46,629,944 | Trans | 2.2 × 10⁻¹⁴⁵ | 4.3 × 10⁻⁵ | MAP kinase-activated protein kinase 3 (MAPKAPK3) |
| 155   | PROZαε, PCID2β, CUL4Aα | rs1515863 | 13  | 113,839,747 | Trans | 2.5 × 10⁻¹⁰⁶ | 6.2 × 10⁻⁹ | Dual specificity mitogen-activated protein kinase 2 (MAP2K2) |
| 67    | DHX39λ, TXNL4β, PMFBP1α, HPRβ, DHDH5α, HPβ | rs9302635 | 16  | 72,144,174 | Cis | 6.3 × 10⁻⁵⁴ | 6.2 × 10⁻¹⁹ | Haptoglobin (HP) |
| 82    | SARMIαβε, VTNα, SLC46A1α, TMEM99β, ROLDIP2α, TMEM97β | rs2239908 | 17  | 26,725,265 | Trans | 2.8 × 10⁻¹⁰ | 1.4 × 10⁻⁸ | Ferritin (FT1H FTL) |
|      |         |         |     |         | Trans | 9.5 × 10⁻⁴⁸ | 3.1 × 10⁻¹⁰ | Semaphorin-3 A (SEMA3A) |
| 54    | APOCαλ, APOC4-APOC2α, APOCα, APOE, APOC1ε,... | rs5167 | 19  | 45,448,465 | Trans | 3.0 × 10⁻²⁴ | 3.9 × 10⁻⁸ | Calcium/calmodulin-dependent protein kinase type 1D (CAMKID) |
| 106   | TRPC4Aαμβε, EDÉM2αβε, PROCRα, GSSββ, MYH7βαβ,... | rs867186 | 20  | 33,764,554 | Trans | 2.1 × 10⁻⁶⁹ | 9.5 × 10⁻⁶ | WNT1-inducible-signaling pathway protein 1 (WISP1) |
|      |         |         |     |         | Trans | 5.5 × 10⁻³⁸ | 9.0 × 10⁻²⁴ | Granulocyte colony-stimulating factor (CSF3) |
|      |         |         |     |         | Trans | 1.1 × 10⁻¹⁰ | 7.7 × 10⁻⁵ | Vitamin K-dependent protein C (PROC) |

Table 2 | List of replicated trans-pQTLs

Loci that comprise at least one replicated trans-association. Loci are referenced in this study by numbers ranging from 1 to 451 (strongest to weakest) and sorted here by chromosome position. P values are for the association with inverse-normal scaled protein levels from linear regression with genotype; see Supplementary Data 1 for full data of all 539 SNP-protein associations at 451 loci, including statistics for association with alternatively raw- and log-normal-scaled protein levels and estimated replication power. Candidate genes for the protein associations were annotated by considering the following criteria: a variant in linkage disequilibrium with the sentinel SNP (r² > 0.8) is located in the gene transcript (superscript a), a variant hits a regulatory element of that gene (superscript b), a variant is a cis-eQTL (superscript c), a variant is a trans-eQTL (superscript d), a variant is protein changing (superscript e). The list of candidate genes in this table is limited to the five most plausible candidate genes for each locus. The full list is available online and in Supplementary Data 1. Every trans-pQTL implies the existence of a functional and causal link between a cis-encoded candidate gene and the trans-encoded target protein(s).

disease-associated proteins vary in response to changes in the genome. Moreover, partial correlations to functionally related proteins, as reported here, may further substantiate and extend hypotheses generated from pQTLs (see example in Fig. 1c). In the following sections, we show how pQTLs identified in this study reveal new insights into multiple disease-associated pathways identified in previous GWAS.

Auto-immune disorders. Ankylosing spondylitis is a common cause of inflammatory arthritis and affects one in 200 Europeans. Evans et al.37 identified two independent ankylosing spondylitis-risk variants in the endoplasmic reticulum aminopeptidase 1 (ERAP1) gene; they reported that the major allele of rs30187 and the minor allele of rs10050860 were protective. ERAP1 is involved in trimming peptides before HLA class I presentation. It has recently attracted attention as a drug target for auto-immune disorders38. Several studies showed that ERAP1 was present in blood, localized to exosome-like vesicles and present in the extracellular space39. Here we found that the two previously identified ankylosing spondylitis-risk variants were associated,
in an additive manner, with increasing levels of circulating ERAP1 protein (Fig. 4a). Similarly, mRNA sequences isolated from lymphoblastoid cells showed that ERAP1 mRNA expression also increased with increasing numbers of ankylosing spondylitis-risk alleles (Fig. 4b). Previous work concluded that the association between ERAP1 and ankylosing spondylitis was mainly driven by genetic differences in how ERAP1 enzymatic activity shaped the HLA-B27 peptidome\(^\text{40,41}\). However, our findings suggest that the auto-immunogenic effects of different ERAP1 protein variants may in addition be modulated by genotype-dependent regulation of protein expression. This observation may have broader implications for treatment approaches to autoimmune disorders that depend on ERAP1 antigen processing.

**Complement system and haem clearance.** We identified two independent variants (rs10494745 and rs10801582), located in the *complement factor H-related 2/4 (CFHR2/CFHR4)* gene locus. These variants were associated in *cis* with haemopexin (HPX) protein levels (Fig. 2a). Lower CFHR4 expression levels were associated with lower HPX protein levels. HPX binds haem with high affinity and transports it from the plasma to the liver, which prevents the accumulation of oxidative species. The rs10494745 variant is a G→E amino acid substitution in CFHR4. It is an eQTL for CFHR4 expression in liver, as it tags the GTEx rs4915318 variant (GTEx, \(P = 1.6 \times 10^{-7}\)). Imputed data revealed a third, strong, and independent signal on SNP rs61818956 (\(P = 6.6 \times 10^{-4}\)), which is located in an intron in the CFHR2/CFHR4 locus. Conditional analysis showed that all three variants were statistically independent, and together, they explained a surprising 61% of the observed variance in a key protein responsible for oxidative stress reduction. Previous studies reported that CFHR4 interacts with complement component 3 (C3) (ref. 42), and in turn, C3 interacts with haem\(^\text{13}\). Those findings suggest a plausible *trans*-acting pathway that links CFHR4 and HPX. That observation may have important consequences on our understanding of the pathologies involving the classical and alternative complement activation pathways.

**Alzheimer’s disease (AD) and mRNA splicing.** Our findings on the major AD-risk variant, rs4420638, may generate particular medical interest. This variant displayed a *cis*-association with increased levels of apolipoprotein E (isoform E2) (APOE) and a concordant *trans*-association with decreased levels of small

---

**Figure 3 | Genotype-dependent co-associations of the plasma proteome and the plasma N-glycome.** Bee swarm plots of total plasma N-glycans GP19 and GP33 (% of total N-glycan content) as a function of rs3760775 and rs8283 genotype, respectively, see inset for glycan structure, Blue squares: N-acetylgalactosamine, green circles: mannose, yellow circles: galactose, purple diamonds: N-acetylneuraminic acid; Scatter plots of total plasma N-glycans GP19 and GP33 as a function of *Complement factor 4 (C4)* and *Galactoside 3(4)-L-fucosyltransferase (FUT3)* genotype (raw data), respectively (b,d), black: major allele homozygotes, red: heterozygotes, green: minor allele homozygotes. Large circles indicate means by genotype. \(P\) values are for the association of glycans with genotype (a,c) and of glycans with protein levels (b,d). \(P\) values are uncorrected from linear regression. The major allele variant of SNP rs3760775 was reported to be associated with the cancer antigen 19-9 and that of SNP rs8283 with increased risk of rheumatoid arthritis.
nuclear ribonucleoprotein F (SNRPF aka RUXF). This dual association was further supported by the finding that the ratio between APOE and SNRPF strengthened the association with rs4420638 by 16 orders of magnitude (p-gain statistic 44). Although this association lacked sufficient replication power, both APOE and SNRPF associations remained nominally significant in the QMDb cohort (P < 0.012), with concordant trends. SNRPF is a core component of U1, U2, U4 and U5 small nuclear ribonucleoproteins, which are the building blocks of the spliceosome. Recently, a knock-down of U1-70 K or inhibition of U1 small nuclear ribonucleoprotein components was shown to increase the levels of amyloid precursor protein 45. An association between this major AD-risk variant and a protein of the spliceosomal machinery has not been reported previously. Taken together, these observations support the implication that protein splicing may be an important factor in AD. Furthermore, the trans-association between SNP rs4420638 and SNRPF had opposite directionality compared with that with APOE levels. These observations suggest that a regulatory mediator is most likely involved. Theoretically, pharmacological targeting of this mediator could cause an increase in SNRPF, which would increase splicing, and potentially decrease amyloid precursor protein levels.

Pharmacogenetics. A pQTL that implicates a drug target may affect patient response to treatment. For example, we observed a strong, replicated cis-association between rs489286 and reduced blood levels of the signalling lymphocytic activation molecule-F7 (SLAMF7) in carriers of the minor allele. This was further confirmed with a SLAMF7-eQTL that showed identical directionality in mRNA sequencing data from lymphoblastoid cell lines (Supplementary Fig. 4). Furthermore, imputed data identified a strong association between SLAMF7 protein levels and a SNP in the SLAMF7 intron, rs11804305, which tags an independent and causative gene variant. We cannot exclude the possibility that rs17313034 affects the SLAMF7 epitope, which could potentially alter SLAMF7-aptamer binding. However, such an alteration might then also affect Elotuzumab binding. This hypothesis can be tested retrospectively in the phase-3 clinical trial cohort.

Drug target validation. Plenge et al.48 suggested that naturally occurring genetic variance could be used to validate drug targets, based on genotype–phenotype dose–response curves. For example, IL6R was previously proposed as a target for preventing coronary heart disease. Recently, tocilizumab, a humanized antibody that targets IL6R, was developed and is currently approved for treating rheumatoid arthritis. Plenge et al.48 required that, for validating drug targets, the gene must include multiple causative variants of known biological function. A large IL6R Mendelian randomization analysis found that SNP rs7529229 was associated with increased IL6R, reduced CRP, reduced fibrinogen, and reduced odds of coronary heart disease (P = 1.53 × 10⁻⁸). The CardiogramPlus consortium GWAS data confirmed that association (P = 1.66 × 10⁻⁸). In the present study, we replicated the IL6R-pQTL (rs4129267, r² = 0.94 with rs7529229). Furthermore, we identified a second SNP in IL6R, rs11804305, which tags an independent and causative variant, since rs4129267 was already established as functional. Therefore, these two SNPs could be used to investigate dose–response curves in the huge data set available from the IL6R-Mendelian randomization consortium. A similar approach is now feasible for all other protein drug targets that were associated here with multiple pQTLs (for examples see Supplementary Fig. 2). What is more, the aptamers that were used in this study to target these proteins can readily be used as intermediate readouts in assessing drug responses and in optimizing the efficacy of lead components.

Application to disease GWAS. Genetic associations with intermediate traits are generally much stronger than associations with disease end points, due to their proximity to the causative variant,
as shown in our previous GWAS with metabolic traits50. Therefore, pQTLs can serve as proxies to fine-map genetic disease associations. This approach can be used to identify potentially causative genes and additional independent genetic signals at a locus identified in a disease-GWAS. In particular, pQTLs can be used to identify true positive associations among associations that do not reach genome-wide significance. For instance, rs12146727 was associated with cardiovascular disease (CVD; \( P = 6.18 \times 10^{-5} \)) in CardiogramPlus and with AD (\( P = 3.10 \times 10^{-5} \)) in the discovery cohort (st1), and \( P = 8.27 \times 10^{-6} \) in a combined analysis of a larger cohort (st12comb)) in the International Genomics of Alzheimer’s Project. A true positive association requires that the association signal must be strengthened with increasing sample numbers, which was the case for the st1 and st12comb cohorts. In the present study, rs12146727 was replicated as a cis-pQTL for complement C1r subcomponent (C1R), and it was replicated as a trans-pQTL for complement C1q subcomponent (C1QA/C1QB/C1QC). These associations support growing evidence that suggests that the complement system has a role in both AD and CVD31. This example also shows how variants that associate with a disease-end point could generate new hypotheses about the role of the co-associated proteins in the disease aetiology. Note that even when these variants have small effect sizes or odds ratios, the related pathways may show large responses to pharmaceutical alteration of these proteins. Hypotheses generated with this approach can be directly tested in animal models, and the existing aptamers can be potentially used as intermediate functional readouts in the drug development process.

Discussion

Genetic studies with yeast52 and lymphoblastoid cell lines53–55 have indicated that cellular protein levels are under strong genetic control. This control was confirmed by the discovery of many cis-acting genetic variants in the human blood proteome4–9. Here we described the first large-scale proteomics GWAS on blood plasma proteins derived from a human population. This GWAS represented over 1.1 million individual aptamer binding experiments. By design, our panel of more than 1,100 aptamer targets was highly enriched in biomedically relevant blood circulating proteins, which was reflected in the large overlap of pQTLs with risk loci identified in disease-GWAS. The data generated from this study can be used in future investigations to identify and validate causative variants identified in disease GWAS. For instance, aptamer-based readouts can be used as intermediate traits in CrispR-based experiments, as exemplified with the FTO-obesity association described by Claussnitzer et al.56,57.

Surprisingly, we found that up to 60% of the naturally occurring variance in the blood plasma levels of essential proteins could be explained by two or more independent variants of a single gene located on another chromosome (trans-associations). We identified strong genetic associations with intermediate, and most likely functional, traits related to proteins involved in complex disorders. While many of our associations connect these proteins to disease pathways through shared associations, it should be borne in mind that confounding is always a possibility and needs to be ruled out by further experimentation. Our data can be used in future studies to establish dose-response curves for drug-target validation49. A greater understanding of the genetic control of circulating levels of protein drug targets and biomarkers may improve pharmaceutical interventions and clinical trials. Because all the aptamers used in this study were synthetically generated and are well defined, they can be readily developed into specific assays for precise clinical applications. For instance, the SLAMF7-binding aptamer identified here, which revealed a potential genetic effect on Elotuzumab, may be developed directly into a clinical assay for identifying differential responders to immunotherapy. Moreover, this concept can be generalized to other drug targets and biomarkers.

Despite the variable baseline conditions among the participants of the replication cohort (not fasting, high prevalence of diabetes and multiple ethnic backgrounds), we replicated 82% of all sufficiently powered associations. This result emphasized the robustness of the replicated associations. It also suggests that many of the Bonferroni-significant associations that could not be replicated in our QMDiab cohort may be replicated in future studies. About 20% of all assayed proteins had a significant pQTL association. This number is expected to increase in future more highly powered studies.

Genetic variance in a protein sequence may affect its higher order structure, and thus, its aptamer-binding affinity. Similarly, alterations in protein structure may affect binding efficacy in immunoblot-based methods and peak mass in targeted mass spectrometry (MS)-based methods. These issues remain to be addressed. In this study, we showed that structural-based epitope effects on a particular pQTL may be identified or ruled out in several ways, including: allele-specific transcription analysis (for example, CPNEN1, Supplementary Fig. 5), co-associated eQTLs, particularly when they include multiple genetic variants (for example, ERAPI, Fig. 4), the absence of correlated SNPs that alter the protein structure (for example, based on SNIPA-annotated 1000 Genomes data), and replication on different platforms (Supplementary Tables 2, 4 and 5). We did not directly follow up on our results using mass spectroscopy. However, Ngo et al.38 recently showed that response curves for selected aptamer-enriched proteins were linear over a wide dynamic range of spiked protein concentrations. Most importantly, trans-associations, which were a central focus of this study, are not affected by this type of potential artefact.

In summary, our GWAS demonstrated the power of linking the genome to disease end points via the blood proteome. As we have shown in the examples, mining of our data can reveal a plethora of new insights into biological processes and provide a wealth of functional information that is beyond the focus of a single publication. Therefore, we have provided additional interpretations of selected loci in Supplementary Note 1. Furthermore, our complete results are freely accessible for further analyses of pQTL associations with past and future disease GWAS, on an integrated web-server at http://proteomics.gwas.eu.

Methods

Study population. The KORA F4 study is a population-based cohort of 4,261 participants living in southern Germany. Study participants were recruited between 2006 and 2008 comprising individuals who, at that time, were aged 32–81 years. KORA F4 is the follow-up study of the KORA S4 survey conducted in 1999–2001 (4,261 participants). The study design, standardized sampling methods and data collection (medical history, questionnaires, anthropometric measurements) have been described in detail elsewhere (ref. 11 and references therein). For this study, 1,000 individuals were randomly selected from a subset of 1,800 already deeply phenotyped KORA F4 study participants13,59,60. All study participants have given written informed consent and the study was approved by the Ethics Committee of the Bavarian Medical Association.

The QMDiab is a cross-sectional case-control study that was conducted between February and June 2012 at the Dermatology Department of HMC in Doha, Qatar. QMDiab has been described previously and comprises male and female participants in near equal proportions, aged between 23 and 71 years, mainly from Arab, South Asian and Filipino descent12. The initial study was approved by the Institutional Review Boards of HMC and Weill Cornell Medicine—Qatar (WCM-Q) (research protocol #11131/11). Written informed consent was obtained from all participants.
Blood sampling. Blood samples for omics-analyses and DNA extraction were collected between 2006 and 2008 as part of the KORA F4 follow-up. To avoid variance due to circadian rhythm, blood was drawn in the morning between 08:00 and 10:30 after a period of at least 10 hours overnight fasting. Blood was collected without stasis and was kept at 4 °C until centrifugation. The material was then centrifuged for 10 min (2,750 × g at 15 °C). Plasma samples were aliquoted and stored at −80 °C until assayed on the SOMAscan platform. For QMDiab, non-fasting plasma specimens were collected in the afternoon, after the general operating hours of the morning clinic, and processed using standardized protocols. Cases and controls were collected as they became available, in a random pattern and at the same location using identical protocols, instruments and study personnel. Samples from cases and controls were processed in the lab blinded to their identity. After collection, the samples were stored on ice for transportation to WCM-Q. Within six hours after sample collection all samples were centrifuged at 2,500g for 10 min, aliquoted, and stored at −80 °C until analysis.

Genotyping. The Affymetrix Axiom Array was used to genotype 3,788 participants of the KORA S4 study. After thorough quality control (total genotyping rate was 99.8%) and filtering for minor allele frequency > 1%, a total of 309,946 autosomal SNPs was kept for the GWAS analysis. One-thousand independent association loci, we lumped in a first step for every given probe all SNPs on a given probe and defined genetically independent associations as those variants that differed in their order and position. To define genetically independent association loci, we lumped in a first step for every given probe all SNPs on a given probe and defined genetically independent associations as those variants that differed in their order and position. To define genetically independent association loci, we lumped in a first step for every given probe all SNPs on a given probe and defined genetically independent associations as those variants that differed in their order and position.

Proteomics measurements. For KORA, the SOMAscan platform was used to quantify protein levels. It has been described in detail before 5,10,61–65. Briefly, 1,231,345 were autosomal variants. The total genotyping rate of these remaining variants was 99.7%. DNA was extracted from 359 samples from QMDiab and genotyped by the WCM-Q genomics core facility using the Illumina Omni 2.5 array (version 8). High-quality genotype data of 2,338,671 variants was obtained for 353 samples; data for six samples was excluded due to a low overall call rate (<90%). Removal of duplicate variants left 2,352,367 variants. In all, 134,830 variants were included due to complete genotype data (PLINK option --log-odds 0.03), leaving 2,192,532 variants. In all, 941,058 variants were removed due to minor allele threshold (PLINK option --maf 0.05), leaving 1,251,474 variants. In all, 28,175 variants were removed due to violation of Hardy-Weinberg equilibrium (PLINK option --hwe 1E-6), leaving 1,223,299 variants. Of these variants, 1,221,345 were autosomal variants. The total genotyping rate of these remaining variants was 99.7%.

Replication. Using the QMDiab proteomics data as dependent variables, linear regression models were fitted using PLINK (version 1.90, version b3), with age, sex, center and probe index as independent variables. The first three principal components separate the three major ethnicities and the three proteomics principal components account for variability introduced by a low degree of cell haemolysis. Genomic inflation was low (λ genomic = 1.026, λ probe = 1.116; no correction was applied). A tag SNP for replication of each of the 451 Bonferroni significant loci was selected using the strongest association among all imputed SNPs in the discovery study which had a correlation of r² > 0.8 with the sentinel SNP. In all, 387 tag SNPs for the 451 originally identified SNPs could be identified for replication, covering 462 SNP-probe pairs out of the originally identified 539 SNP-probe pairs. To consider an association as replicated, we, therefore, required P < 1.08 × 10⁻⁴ (0.05/462), 384 (83.1%) displayed nominal significance (P < 0.05). Forty-eight of these associations had a discordant trend. We estimated the statistical power for the replication by sampling. For each association we randomly selected without replacement 338 individuals from the KORA cohort and computed the P value for a linear regression with genotype on that subset. We repeated this 100 times and reported the 95th largest P value from this empirical distribution as the P value that can be expected to be obtained with 95% power (p95). Based on this analysis, we found that 208 SNP-probe pairs with a suitable tag SNP in QMDiab had 95% replication power (p95 < 1.08 × 10⁻⁷). 171 of these 208 (82.2%) fully replicated in QMDiab, 198 (95.2%) displayed at least nominal significance.

Replication of previous SOMAscan based cis-associations. Using the SOMAscan platform on 100 samples, Lourdusamy et al. 2007 reported 60 cis-associations at a false discovery rate of 5% chance. Forty-eight of these associations had a significant tag SNP in the genotyped KORA data set (r² > 0.5) and confirmed the association on an imputed SNP that allowed for replication. Thirty-four of these SNPs were replicated (P < 0.05/120, conservatively accounting for replication attempts on tagged and imputed SNPs). The first non-replicated association had link 21 (Supplementary Table 3).

Replication of immunoassay based cis-associations. Kim et al. 2007 reported 28 cis-associations for 27 analytes in the ADNI cohort using plasma proteomic data by multiplex immunoassay on the Myriad Rules Based Medicine (RB) Human DiscoveryMAP panel v1.0 on the Lumine × 100 platform. Out of 17 associations that had overlapping probes, thirteen were replicated (P < 0.05/17) (Supplementary Table 4). Melzer et al. 2010 tested 40 protein levels determined by immunoassay and reported 8 pQTLs. We replicated their second strongest association with IL6R. Their strongest association is at the ABO locus with TNFα. However, TNFα is not on the SOMAscan panel. We also found numerous other strong associations at the ABO locus. Two weak associations of Melzer et al. (SIB2 and CRP) were not significant in our study. The remaining associations from Melzer et al. were not targeted by our panel. Enroth et al. 2006 used a multiplexed immunoassay and reported 23 cis-pQTLs. Two of the proteins (four associations) targeted by their assay were not covered by our panel (VEGF-D, Ep-CAM) and one was located on the X-chromosome and not tested for cis-association here (CD40-1). Of the remaining 18 cis-pQTLs, twelve associations were replicated (P < 0.05/18) (Supplementary Table 5).

Replication of MS-based cis-associations. Johansson et al. 2016 reported 5 associations and we found cis-pQTLs for four of them (Haptoglobin (HP), Alpha-1-antitrypsin
isoforms were targeted by our panel. Liu et al. reported protein associations using MS in lymphoblastoid cell lines. However, this study was targeting intracellular proteins and had no overlap with blood circulating proteins targeted by our panel and could, therefore, not be used for replication.

Proteome annotation. We used SOMAmer probe identifiers as primary protein identifiers (Supplementary Table 1). Some SOMAmer targets map to multiple Uniprot identifiers (N = 41). They either refer to protein complexes (N = 32) that are encoded at multiple genome loci, or to different variants of a same protein encoded at a single gene locus (N = 9). In all, 1,090 SOMAmer targets were encoded on autosomal chromosomes, 30 targets were encoded on the X-chromosome, and four targeted viral proteins. Genome positions for all SOMAmer targets were retrieved from Ensembl (http://www.ensembl.org) using probe-specific Uniprot identifiers provided by SomaLogic. We used Ingenuity Pathway Analysis (IPA) (http://www.ingenuity.com) to retrieve additional information related to each probe. IPA provides a rich expert-curated knowledgebase of literature-based protein-related information and requires unique mapping to protein identifiers. Forty-one probes were present that target proteins with multiple Uniprot IDs, and these were excluded from the IPA analysis. A further 29 probes were excluded because multiple probes target a same protein. Eight Uniprot IDs could not be mapped by IPA (LG1-1, LD78-beta, NKG2D, HSP 70, and four viral proteins). Ultimately, 1,045 probes with unique Uniprot IDs remained for the IPA annotation.

Functional annotation of the associations. We used the SNiPA server (v3.1, http://snipa.org) to annotate 435 out of our 451 lead SNPs (Supplementary Fig. 9). Sixteen SNPs were not available in the SNiPA database (poly-allelic variants are currently excluded). SNiPA provides annotations for all SNPs that are in linkage disequilibrium (R² > 0.8) with and no more than 500 kb distant from a sentinel SNP using genome assembly data based on GRCh37.p13, Ensembl version 82, 1000-Genomes (phase 3, version 5) data, and GTEx (release 4) eQTL associations for 13 tissues (see columns SNiPA_... in Supplementary Data 1). SNiPA also provides primary effect predictions using the Ensembl VEP tool53, and all GWAS association data from the GWAS catalogue49, metabolite associations (http://gwas.eu) and dbGaP (columns GWAS_TRAITS, mGWAS_TRAITS, dbGaP_TRAITS, accessed October 2015). Updates can be retrieved online at http://snipa.org using the block-annotation tool.

Methylation GWAS. Data was downloaded from the BIOS QTL browser69 http://genenet.nl/biosqtlbrowser (accessed 9 Feb. 2016). Tagging SNPs (r² > 0.8) were identified using SNP-SNP correlations from KORA imputed genotypes and meQTLs with BIOS-P values < 10^-9 were retrieved.

Metabolomics GWAS. GWAS associations with mQTLs were obtained using the GWAS-server (http://gwas.eu). This web-server provides access to raw association data from the studies by Suhr et al.49, Shin et al.20 and Raffler et al.76. Extracted associations were limited to mQTL-P values < 10^-8 and further requiring P-gain > 10^3 for ratios44 (Supplementary Table 8).

Disease GWAS lookup. We used SNiPA to identify overlapping disease-GWAS entries. We further downloaded publicly available association data for 70 clinically relevant traits from the web-sites of 14 large disease GWAS consortia, based on a list of available GWAS, and downloaded from https://www.med.unc.edu/pgc/ downloads. Supplementary Data 4 provides a list with links to all downloaded files and the corresponding publications.

mRNA sequencing and allele specific transcription analysis. Lappalainen et al.27 report mRNA sequencing of 462 lymphoblastoid cell lines of the 1,000 Genomes Project. RNA sequencing data was downloaded from EBI (http://www.ebi.ac.uk/Tools/teuvadis-das/) Analysis was limited to 373 samples of European ancestry. We used CLCbio genomics workbench (Qiagen Inc.) to align reads, calculate RPKM (Reads Per Kilobase of transcript per Million mapped reads) values and analyse the data.

Gaussian Graphical Network (GGM) construction. Using raw (unscaled) data, regressing out age = gender + bmi (997 samples with data for 1,124 variables) we computed a GGM using the ggm.estimate.pcor function from the R GneNet package. The estimated optimal shrinkage intensity lambda (correlation matrix) was 0.187. We obtained 3,943 GGM edges connecting 1,092 protein nodes with Bonferroni significant partial correlation coefficients (P < 7.9 x 10^-9 = 0.05/(1,124*1,123/2)), provided as Supplementary Data 3. We further performed SNP-probe association analyses using all associations reported in the GWAS catalogue (identified using SNiPA at linkage disequilibrium r² > 0.8). We also added all SNP-disease associations from 84 clinically relevant traits of 14 large GWAS consortia that had a GWAS-P value P < 10^-8. These SNP-disease edges were further manually curated to ascertain unique SNP-disease pairs. SNP association to disease-related protein levels were excluded.

Data availability. All summary statistics and association data are freely available (Supplementary Data 5), accessible online on an integrated web-server at http://proteomics.gwas.eu. A fully functional docker-based version of the web-server can also be freely downloaded from this link for local installation and network-free usage (HTML5-based, no extra software is required). The informed consent given by the study participants does not cover posting of participant level phenotype and genotype data in public databases. However, data for KORA are available upon request from KORA-gen (http://epi.helmholtz-muenchen.de/kora-gen). Requests are submitted online and are subject to approval by the KORA board.

References
1. Suhre, K. et al. Human metabolic individuality in biomedical and pharmaceutical research. Nature 477, 54–60 (2011).
2. Stunnenberg, H. G. & Hubner, N. C. Genomics meets proteomics: identifying the culprits in disease. Hum. Genet. 133, 689–700 (2014).
3. Gutierrez-Arcelus, M., Rich, S. S. & Raychaudhuri, S. Autoimmune diseases—connecting risk alleles with molecular traits of the immune system. Nat. Rev. Genet. 17, 160–174 (2016).
4. Melzer, D. et al. A genome-wide association study identifies protein quantitative trait loci (pQTLs). PLoS Genet. 4, e1000072 (2008).
5. Louridas, A. et al. Identification of cis-regulatory variation influencing protein abundance levels in human plasma. Hum. Mol. Genet. 21, 3719–3726 (2012).
6. Johansson, Å. et al. Identification of genetic variants influencing the human plasma proteome. Proc. Natl Acad. Sci. USA 110, 4673–4678 (2013).
7. Kim, S. et al. Influence of genetic variation on plasma protein levels in older adults using a multi-analyte panel. PLoS ONE 8, e70269 (2013).
8. Liu, Y. et al. Quantitative variability of 342 plasma proteins in a human twin population. Mol. Syst. Biol. 11, 786 (2015).
9. Enroth, S., Johansson, A., Enroth, S. B. & Gyllensten, U. Strong effects of genetic and lifestyle factors on biomarker variation and use of personalized cutoffs. Nat. Commun. 5, 4684 (2014).
10. Gold, L. et al. Apteramer-based multiplexed proteomic technology for biomarker discovery. PLoS ONE 5, e15002 (2010).
11. Wichmann, H.-E., Gieger, C. & Illig, T. KORA-gen—resource for population genetics, controls and a broad spectrum of disease phenotypes. Gesundheitswesen 67, S26–S30 (2005).
12. Mook-Kanamori, D. O. et al. 1,5-Anhydroglucitol in saliva is a noninvasive marker of short-term glycemic control. J. Clin. Endocrinol. Metab. 99, E479–E483 (2014).
13. Arnold, M., Raffler, J., Pfeifer, A., Suhre, K. & Kastenmüller, G. SNP_A: an interactive, genetic variant-centered annotation browser. Bioinformatics 31, 1334–1336 (2014).
outside the catalytic and peptide-binding sites on the processing of natural HLA-B27 ligands. J. Biol. Chem. 289, 3978–3990 (2014).

41. Edwards, E., Edwards, C. J., Elliott, T. & James, E. Naturally occurring erpal haplotypes encode functionally distinct alleles with fine substrate specificity. J. Immunol. 191, 35–43 (2013).

42. Hellwage, J. et al. Functional properties of complement factor H-related proteins FHR-3 and FHR-4: binding to the C3d region of C3b and differential regulation by heparin. FEBS Lett. 462, 345–352 (1999).

43. Frimat, M. et al. Complement activation by heme as a secondary hit for atypical hemolytic uremic syndrome complement activation by heme as a secondary hit for atypical hemolytic uremic syndrome. Blood 122, 282–292 (2013).

44. Petersen, A.-K. et al. On the hypothesis-free testing of metabolite ratios in genome-wide and metabolome-wide association studies. BMC Bioinformatics 17, 20 (2012).

45. Bui, B. et al. U1 small nuclear ribonucleoprotein complex and RNA splicing alterations in Alzheimer’s disease. Proc. Natl Acad. Sci. USA 110, 16562–16567 (2013).

46. Miura, K. et al. Genome-wide association study of HPV-associated cervical cancer in Japanese women. J. Med. Virol. 1158, 1153–1158 (2014).

47. Lional, S. et al. Elotuzumab therapy for relapsed or refractory multiple myeloma. N. Engl. J. Med. 373, 621–631 (2015).

48. Plenge, R. M., Scollnick, E. M. & Altshuler, D. Validating therapeutic targets through human genetics. Nat. Rev. Drug Discov. 12, 581–594 (2013).

49. Interleukin-6 Receptor Mendelian Randomisation Analysis (IL6R MR Consortium. The interleukin-6 receptor as a target for prevention of coronary heart disease: a mendelian randomisation analysis. Lancet 379, 1214–1224 (2012).

50. Suhre, K. & Gieger, C. Genetic variation in metabolic phenotypes: study designs and applications. Nat. Rev. Genet. 13, 759–769 (2012).

51. Shen, Y., Yang, L. & Li, R. What does complement do in Alzheimer’s disease? Old molecules with new insights. Transl. Neurodegener. 2, 21 (2013).

52. Foss, E. I. et al. Genetic basis of proteome variation in yeast. Nat. Genet. 39, 1369–1375 (2007).

53. Wu, L. et al. Variation and genetic control of protein abundance in humans. Nature 499, 79–82 (2013).

54. Hause, R. J. et al. Identification and validation of genetic variants that influence transcription factor and cell signaling protein levels. Am. J. Hum. Genet. 95, 1571–1584 (2014).

55. Garge, N. et al. Identification of quantitative trait loci underlying proteome variation in human lymphoblastoid cells. Mol. Cell. Proteomics 9, 1383–1399 (2010).

56. Claussnitzer, M. et al. Leveraging cross-species transcription factor binding site patterns: From diabetes risk loci to disease mechanisms. Cell 156, 343–358 (2014).

57. Claussnitzer, M. et al. FTO obesity variant circuitry and adipocyte browning in humans. N. Engl. J. Med. 373, 895–907 (2015).

58. Ngo, D. et al. Apteraxin-based proteomic profiling reveals novel candidate biomarkers and pathways in cardiovascular disease. Circulation 134, 270–285 (2016).

59. Illeg, T. et al. A genome-wide perspective of genetic variation in human metabolism. Nat. Genet. 42, 137–141 (2010).

60. Petersen, A. -K. K. et al. Epigenetics meets metabolomics: an epigenome-wide association study with blood metabolic traits. Hum. Mol. Genet. 23, 534–545 (2014).

61. Kraemer, S. et al. From SOMAmer-based biomarker discovery to diagnostic and clinical applications: a SOMAmer-based, streamlined multiplex proteome assay. PLoS ONE 6, e26332 (2011).

62. Hathout, Y. et al. Large-scale serum protein biomarker discovery in Duchenne muscular dystrophy. Proc. Natl Acad. Sci. 112, 201507719 (2015).

63. Sattlercker, M. et al. Alzheimer’s disease biomarker discovery using SOMAscan multiplexed protein technology. Alzheimers Dement. 10, 724–734 (2014).

64. Kudde, S. J. et al. Candidate blood proteome markers of Alzheimer’s disease onset and progression: a systematic review and replication study. J. Alzheimers Dis. 38, 515–531 (2014).

65. Menni, C. et al. Circulating proteomic signatures of chronological age. J. Gerontol. Ser. B. Biol. Sci. Med. 70, 809–816 (2015).

66. Chang, C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4, 7 (2015).

67. McLaren, W. et al. Deriving the consequences of genomic variants with the ensemble API and SNP effect predictor. Bioinformatics 26, 2069–2070 (2010).

68. Hindorff, L. A. et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl Acad. Sci. USA 106, 9362–9367 (2009).

69. Bonder, M. J. et al. Disease variants alter transcription factor levels and methylation of their binding sites. Nat. Genet. 49, 113–118 (2017).
70. Raffler, J. et al. Genome-wide association study with targeted and non-targeted NMR metabolomics identifies 15 novel loci of urinary human metabolic individuality. *PLoS Genet.* **11**, e1005487 (2015).

71. Trbojević Akmacić, I. et al. High throughput glycomics: optimization of sample preparation. *Biochemistry* **80**, 934–942 (2015).

72. Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).

**Acknowledgements**

This work was supported by ‘Biomedical Research Program’ funds at Weill Cornell Medicine in Qatar, a program funded by the Qatar Foundation. The statements made herein are solely the responsibility of the authors. M. Arnold was supported by the Helmholtz cross-program topic ‘Metabolic Dysfunction’. D. Mook-Kanamori was supported by Dutch Science Organization (ZonMW-VENI Grant 916.14.023). The KORA study was initiated and financed by the Helmholtz Zentrum München—German Research Center for Environmental Health, which is funded by the German Federal Ministry of Education and Research (BMBF) and by the State of Bavaria. Furthermore, KORA research was supported within the Munich Center of Health Sciences (MC-Health), Ludwig-Maximilians-Universität, as part of LMUinnovativ. The KORA Study Group consists of A. Peters (speaker), J. Heinrich, R. Holle, R. Leidl, C. Meisunger, K. Strauch and their co-workers, who are responsible for the design and conduct of the KORA studies. We gratefully acknowledge the contribution of all members of field staff conducting the KORA F4 study. We thank the staff from HMC dermatology department and WCM-Q clinical research core for their contribution to QMDiab. We thank Brian Sellers from SomaLogic for support with measuring the QMDiab samples at WCM-Q. We acknowledge free access to summary statistics provided by the GWAS consortia listed in Supplementary Data 4. We thank Jenine Davidson for the design of Figure 1a. Most of all, we thank all study participants of KORA and QMDiab for their invaluable contributions to this study.

**Author contributions**

Jointly supervised research: K.S., J.G., C.G. Conceived and designed the experiments: K.S., J.G., C.G. Performed the experiments: R.E., J.G., E.K.A., Y.A.M., J.M., H.S., G.L., M.P., Performed statistical analysis: K.S., C.G., A.W. Analysed the data: K.S., A.M.B., R.J.C., J.G., G.T., M.A., C.G., G.K., J.R., A.W. Contributed reagents, materials, or analysis tools: K.S., M.A.S., M.A., H.G., G.K., A.P., J.R., K.S., D.O.M., R.K.D., L.G. Wrote the paper: K.S. All authors discussed the results and reviewed the final manuscript.

**Additional information**

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: M.P., G.L., K.D. and L.G. are working for or have stakes in Genos Ltd. and Somalogic Inc., respectively. The remaining authors declare competing no financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Suhre, K. et al. Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat. Commun.* **8**, 14357 doi: 10.1038/ncomms14357 (2017).

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/
Erratum: Connecting genetic risk to disease end points through the human blood plasma proteome

Karsten Suhre, Matthias Arnold, Aditya Mukund Bhagwat, Richard J. Cotton, Rudolf Engelke, Johannes Raffler, Hina Sarwath, Gaurav Thareja, Annika Wahl, Robert Kirk DeLisle, Larry Gold, Marija Pezer, Gordan Lauc, Mohammed A. El-Din Selim, Dennis O. Mook-Kanamori, Eman K. Al-Dous, Yasmin A. Mohamoud, Joel Malek, Konstantin Strauch, Harald Grallert, Annette Peters, Gabi Kastenmüller, Christian Gieger & Johannes Graumann

Nature Communications 8:14357 doi: 10.1038/ncomms14357 (2017); Published 27 Feb 2017; Updated 11 Apr 2017

The original version of the Supplementary Information attached to this Article did not include Supplementary Note 1. The HTML has now been updated to include a corrected version of the Supplementary Information.