OXFORD

Human Molecular Genetics, 2022, Vol. 00, 00, 1–10
https://doi.org/10.1093/hmg/ddac243
Advance access publication date 28 September 2022
Original Article

Differences and commonalities in the genetic architecture of protein quantitative trait loci in European and Arab populations

Gaurav Thareja[1,2], Aziz Belkadi[1,2], Matthias Arnold[3,4], Omar M.E. Albagha[5,6], Johannes Graumann[7], Frank Schmidt[8], Harald Grallert[9,10,11], Annette Peters[9,11,12,13], Christian Gieger[9,10,11], The Qatar Genome Program Research Consortium[12] and Karsten Suhre[1,2,*]

[1]Bioinformatics Core, Weill Cornell Medicine-Qatar, Education City, 24144 Doha, Qatar
[2]Department of Biophysics and Physiology, Weill Cornell Medicine, NY 10065, New York, USA
[3]Institute of Computational Biology, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstraße 1, Neuherberg 85764, Germany
[4]Department of Psychiatry and Behavioral Sciences, Duke University, NC 27710, USA
[5]College of Health and Life Sciences, Hamad Bin Khalifa University, 34110 Doha, Qatar
[6]Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, EH4 2XU, Edinburgh, UK
[7]Institute of Translational Proteomics, Department of Medicine, Philipps-Universität Marburg, Marburg, Germany
[8]Proteomics Core, Weill Cornell Medicine-Qatar, Education City, 24144 Doha, Qatar
[9]Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstraße 1, Neuherberg 85764, Germany
[10]Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstraße 1, Neuherberg 85764, Germany
[11]German Center for Diabetes Research (DZD), Ingolstädter Landstraße 1, Neuherberg 85764, Germany
[12]German Center for Cardiovascular Research (DZHK), Partner Site Munich Heart Alliance, Munich, Germany
[13]Department of Epidemiology, Institute for Medical Information Processing, Biometry, and Epidemiology, Ludwig-Maximilians-University Munich, 81377 Munich, Germany
[14]A list of consortium authors and affiliations appears at the end of the paper
[*]To whom correspondence should be addressed at: Weill Cornell Medicine-Qatar, Education City, P.O. 24144, Doha, Qatar. Tel: +974.33541843; Email: kas2049@qatar-med.cornell.edu

Abstract

Polygenic scores (PGS) can identify individuals at risk of adverse health events and guide genetics-based personalized medicine. However, it is not clear how well PGS translate between different populations, limiting their application to well-studied ethnicities. Proteins are intermediate traits linking genetic predisposition and environmental factors to disease, with numerous blood circulating protein levels representing functional readouts of disease-related processes. We hypothesized that studying the genetic architecture of a comprehensive set of blood-circulating proteins between a European and an Arab population could shed fresh light on the translatability of PGS to understudied populations. We therefore conducted a genome-wide association study with whole-genome sequencing data using 1301 proteins measured on the SOMAscan aptamer-based affinity proteomics platform in 2935 samples of Qatar Biobank and evaluated the replication of protein quantitative traits (pQTLs) from European studies in an Arab population. Then, we investigated the colocalization of shared pQTL signals between the two populations. Finally, we compared the performance of protein PGS derived from a Caucasian population in a European and an Arab cohort. We found that the majority of shared pQTL signals (81.8%) colocalized between both populations. About one-third of the genetic protein heritability was explained by protein PGS derived from a European cohort, with protein PGS performing ∼20% better in Europeans when compared to Arabs. Our results are relevant for the translation of PGS to non-Caucasian populations, as well as for future efforts to extend genetic research to understudied populations.

Introduction

Genome-wide association studies (GWAS) with complex disease endpoints have revealed many genes and pathways involved in a plethora of pathophysiologies (1). With increasing sample sizes, approaches like Mendelian randomization (MR) allow to evaluate the potential of associated proteins as new drug targets (2), polygenic risk scores (PGS) inform precision medicine and identify individuals who are at elevated risk of developing preventable diseases (3), while genetic correlations provide etiological insights into complex disease traits (4). GWAS with intermediate phenotypes (5), such as gene expression, proteomics and metabolomics complement these studies by providing functional read-outs for disease-relevant pathways (6–8), allowing for drug-target prioritizations and mode of action rationalizations (9).

However, while immensely successful in populations of Northern European ethnicity, and currently catching up in Hispanic, African American, Chinese and Japanese populations, GWAS in many lesser studied populations are still lagging behind (10,11). In

Received: June 10, 2022. **Revised:** September 19, 2022. **Accepted:** September 22, 2022

this context, the question arises as to how far GWAS findings, in particular PGS and MR instruments, may be translated from statistically well-powered studies to under-represented populations.

Previously, we reported a GWAS of 45 clinically relevant traits in the Middle Eastern population of Qatar. We found that PGS derived from studies in Caucasian ethnicities performed poorly when applied to the Qatar population with an average relative performance of 64.7% (s.d. = 15.8%) compared to when applied to Europeans (12). Similarly, a recent GWAS in the UK Biobank (UKB) with 35 blood and urine biomarkers found that PGS performed better in the non-British white UKB sub-population as compared to the South-East Asian and African UKB study participants (13). We reasoned that a trans-ethnic study with a larger number of quantitative traits, such as blood circulating protein levels, may draw a broader picture of the general translatability of PGS between populations, and that this approach may also shed light on the underlying similarities and differences in the genetic architectures of the individual associated with genetic loci.

Proteomics recently became accessible to GWAS, all thanks to the technology advances in affinity proteomics, implemented by the SOMAscan and Olink platforms (7) as well as enhanced mass spectrometry approaches (14), leading to the discovery of pQTLs for hundreds of proteins in studies with several thousand participants (15–20). PGS for the plasma proteome have already been shown capable of assessing the polygenic risk of cardiometabolic diseases (21) and obesity (22), but these studies were restricted to European populations. Here we set out to investigate the genetic architecture of pQTLs in an Arab population, using proteomics measurements that we recently performed in samples from 2935 participants of the Qatar Biobank (QBB) (23,24) using the SOMAscan aptamer-based affinity proteomics platform of Weill Cornell Medicine-Qatar (17).

Our analysis comprises of three parts: first, we conduct a 'classical' GWAS with proteomics in QBB and evaluate the replication of pQTLs from the German KORA and the British INTERVAL studies to identify potential population-specific associations (Fig. 1). Then, we fine-map the pQTL signals of QBB by investigating their colocalization with the pQTLs in INTERVAL (Fig. 2). Finally, we derive protein PGS using summary statistics from the most highly powered protein GWAS to date, the Icelandic deCODE study, and evaluate the performance of these PGS in the KORA and QBB studies by comparing them to directly measured protein levels in two almost equally powered studies of European and Arab ethnicity, respectively (Fig. 3).

Results

A GWAS with 1301 blood circulating protein traits identifies 2685 pQTLs in an Arab population

We conducted a GWAS between 1301 blood circulating protein levels and 10 004 359 genetic variants determined by whole-genome sequencing (WGS) (see methods). We identified 2685 pQTL associations at a Bonferroni level of significance ($P < 3.8 \times 10^{-11} = 5 \times 10^{-8}/1301$), located at 2384 independent genetic loci ($R^2 < 0.1$ between sentinel variants). Of the 1301 proteins assayed in this study, 574 had at least one pQTL in QBB. A total of 1925 (71.7%) of the identified pQTLs were located *in-cis* (382 proteins) with respect to the protein coding gene while 760 (28.3%) were located *in-trans* (282 proteins, Supplementary Material, Table S1). A total of 135 genetic loci (5.7%) had more than one pQTL, 14 loci had five or more pQTLs and the two most pleiotropic loci, *VTN* and C7, had 23 and 36 pQTLs, respectively. Of the 2685 pQTLs, 148 pQTLs (5.5%) had not been previously reported on any variant within a window of +/−10 Mb from the sentinel variant at a significance level of $P < 10^{-8}$ or in linkage disequilibrium (LD) ($R^2 > 0.8$) with a previously reported pQTL at a significance level of $P < 10^{-5}$, nor were they reported in the latest version of SNiPA (25). These pQTLs are considered novel.

Up to 89% of pQTLs identified by European studies replicate in the Arab population

We attempted replication of two major pGWAS: the INTERVAL study reported 1980 associations based on a GWAS with 3301 participants and 3622 protein traits. We replicated 381 out of 617 (61.8%) pQTLs for which we had protein data in QBB and for which a tag variant ($r^2 > 0.8$) was available, using a significance level for replication of $P < 0.05/617$ (Supplementary Material, Table S2). Of 57 pQTLs reported at the pleiotropic *C7* locus, we replicated 26 out of 30 for which we had data, and additionally found a *cis*-pQTL for C7 which had not been reported before. The pleiotropic *VTN* locus had 116 pQTLs in INTERVAL—we replicated 11 out of 18 pQTLs for which we had proteomics data, including the previously reported VTN *cis*-pQTL. The KORA study reported 539 associations based on a GWAS with 997 participants and 1124 protein traits. We replicated 452 out of 508 (89.0%) pQTLs for which protein data and a tag variant ($r^2 > 0.8$) were available, based on a significance level of $P < 0.05/508$ (Supplementary Material, Table S3).

The majority of *cis*-pQTLs are driven by non-protein altering variants, while 6.5% may be due to epitope effects

As we had WGS data available, we could identify all potentially protein altering variants (PAV) that are in high LD with the sentinel variants of the *cis*-pQTLs (LD based on QBB data). Such variants can potentially impact the aptamer binding to the protein epitope without having any other biological effect (Supplementary Material, Table S1). Of the 1925 *cis*-pQTLs, 196 (10.2%) had at least one such PAV in LD, based on a cutoff of $r^2 > 0.8$. For 122 (62.2%) of these pQTLs, the sentinel variant was itself a PAV. However, 70 (35.7%) of the 196 pQTLs with a PAV in LD were also associated with at least one clinical GWAS trait, suggesting that at most 126 of the 1925 pQTLs (6.5%) may be affected by a pure epitope effect without any biomedical consequence. The majority of the *cis*-pQTLs (89.8%) are therefore likely to be driven by regulatory variants.

Locus fine mapping and cross-population colocalization suggest that 81.8% of regions with shared associations between Europeans and Arabs have a similar genetic architecture

We grouped adjacent pQTLs with a distance between lead variants of less than 500 kb into 771 pQTL regions and assigned all variants within a distance of 500 kb from any one of the lead variants to that region. The average number of pQTLs per region was 3.5 and the median was one (Supplementary Material, Table S4). The region with the largest number of pQTLs was at the HLA locus on chromosome 6 in association with MICA protein levels, which had 55 pQTLs in one region plus an additional two regions that were more than 500 kb apart and that were, therefore, treated as independent regions by our approach. We performed region-wise genetic fine-mapping using GCTA COJO with all variants in a region as input (26). The average number of fine-mapped variants per locus was 1.5 with a median of one and the largest number of variants was seven for CD177 antigen and six for MICA (Supplementary Material, Table S4). We then

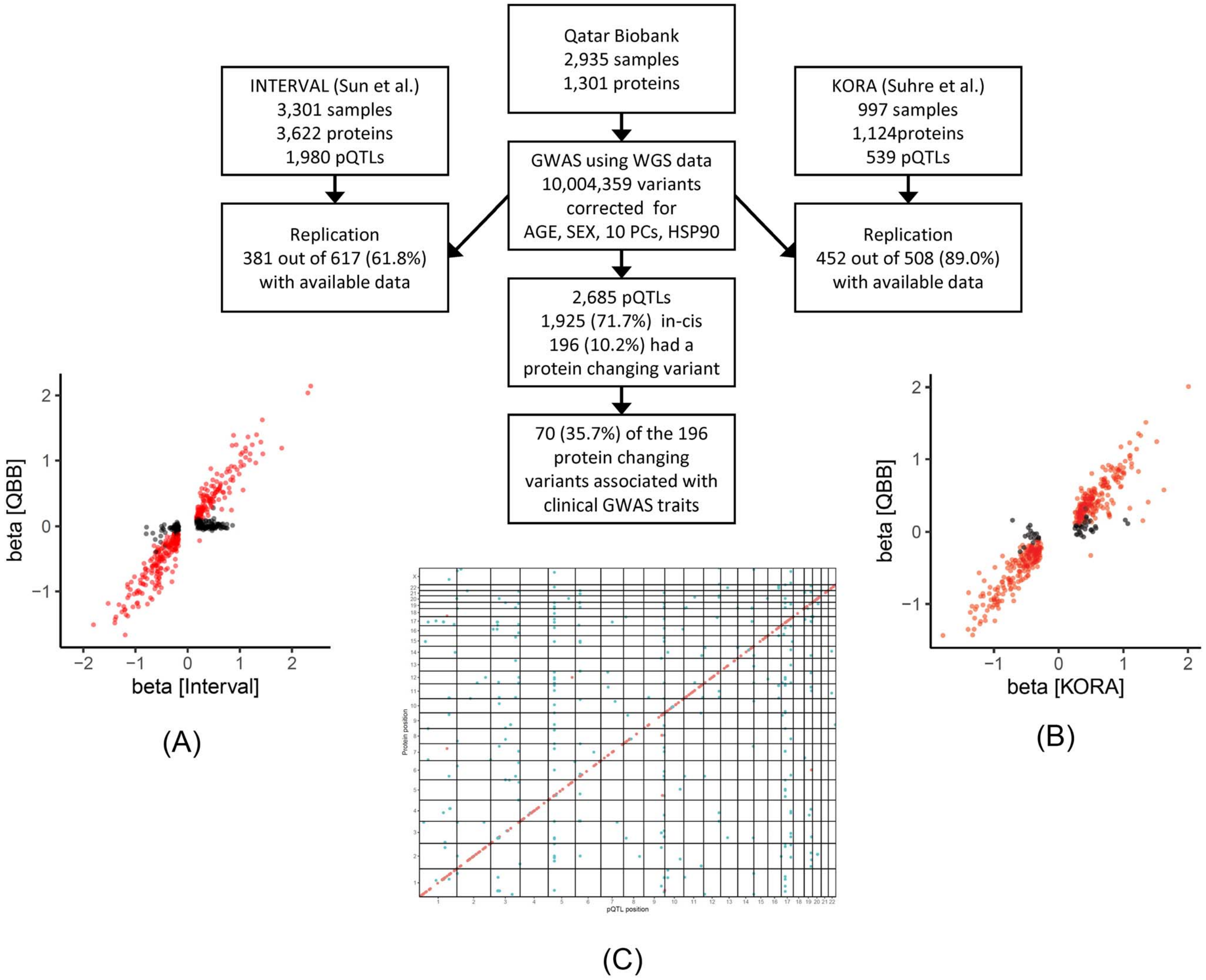

Figure 1. pGWAS study and replication of pQTLs from KORA and INTERVAL. Scatterplot of the effect sizes of all replicable pQTLs between INTERVAL and QBB (**A**) and between KORA and QBB (**B**), 2-D Manhattan plot of all significant QBB pQTLs (**C**); full summary statistics for these replications are in Supplementary Material, Tables S2 and S3; full-size versions and regional association plots for all QBB pQTLs are in Supplementary Material, Figures S1–S3.

compared the genetic architecture of the pQTL regions between QBB and INTERVAL using the approach that is implemented in the coloc package (27) together with SuSiE (Sum of Single Effects) for credible set fine-mapping (28). SuSiE is a novel approach that evaluates the evidence for associations with multiple potential causal variants simultaneously and has been applied to integrate fine-mapping across 148 complex traits in three large-scale biobanks (29). It was also used in a recent plasma proteome analysis in individuals of European and African ancestry (30). Briefly, SuSiE identifies so-called 'credible sets' (CS) of variants that are likely to contain the causal variant and provides corresponding posterior probabilities (PIP) for each of them being possibly causal for the association with the trait. Coloc is usually used to determine whether the association signals of two traits at the same genetic locus are generated by a shared causal variant. This is done using summary statistics from two GWAS conducted in the same population and by analyzing the correlation between the association statistics. In contrast, here we used this approach to ask whether an association signal for the same protein trait correlates between two different populations (Fig. 2). Colocalization was performed between all possible combinations of credible sets. The probabilities for several hypotheses (H0–H4) regarding the question of whether two associations are sharing the same genetic signal or not were then computed. Hypotheses H0, H1 and H2 (none or only one association contains a causal variant) were excluded in our setting, as we only consider regions that contain pQTLs that were significant in both, the INTERVAL and QBB study. Hypothesis H3 states that the two association signals are distinct, and H4 that both associations have the same causal variant. A total of 598 of the 771 regions had shared pQTLs between INTERVAL and QBB and could be further analyzed using this cross-ethnic SuSiE-coloc approach. Of these 598 shared regions, 133 had no significant variant ($P < 5 \times 10^{-5}$) in INTERVAL, 36 had no credible set >95% in INTERVAL, and five had no credible set >95% in QBB and were not further considered. A total of 424 regions with shared pQTL signals could hence be analyzed, out of which 347 regions had at least one H4 (81.8%). The total number of signals that support hypothesis H4 was 454 and the total number of regions that only support hypothesis H3, but no H4, was 77. We provide detailed colocalization plots for all analyzed regions as a resource for future investigations of specific loci (Supplementary Material, Fig. S4).

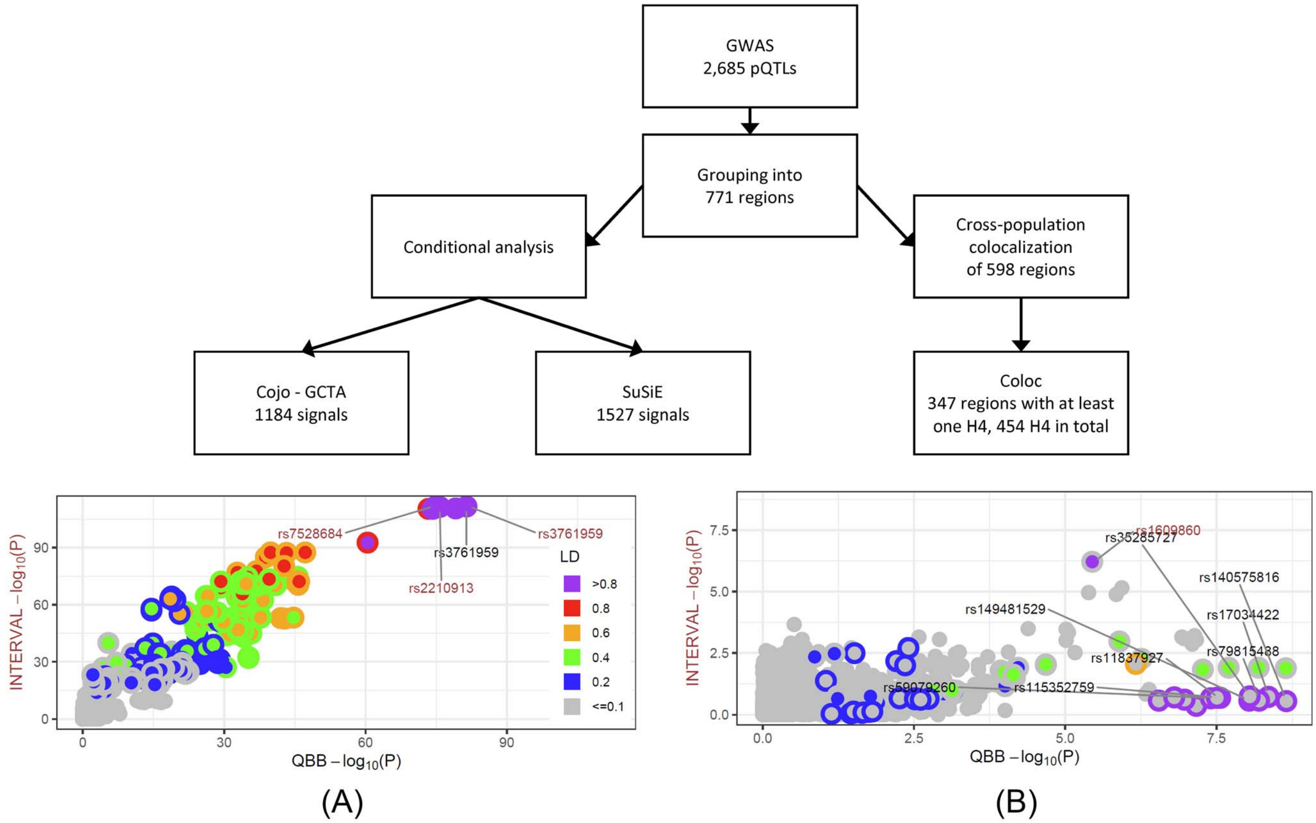

Figure 2. Locus fine mapping and cross-population colocalization analysis. Examples of scatterplots of *P*-values for two regions, one with shared genetic architecture between INTERVAL and QBB (FCRL3, **A**) and one with two unrelated signals (DKK3, **B**); circle colors reflect LD r^2 with the respective lead variant, the circle fill color is for Interval, circle border color represents LD in QBB (see Supplementary Material, Fig. S4 for further detail and Supplementary Material, Table S4 for Cojo, SuSiE and Coloc statistics).

Polygenic scores of protein traits are driven by large effect variants at current study sizes

We generated PGS for 842 proteins that were measured on all versions of the SOMAscan platform in the KORA, QBB and deCODE studies, using the deCODE protein association summary statistics by clumping correlated variants based on LD (r^2_{PGS}) and discarding all associations below a fixed *P*-value cut-off (p_{PGS}). We limited the PGS variant selection to variants that were shared between all three studies ($N = 5\,179\,443$) so that our comparison would not be biased by differences in the availability of variants between the studies (Fig. 3 and Supplementary Material, Table S5). We explored a range of *P*-value cutoffs and r^2 clumping values and evaluated the performance of the PGS by correlating the predicted and measured protein levels using 944 samples of the KORA study and 1155 samples from unrelated participants of QBB (Supplementary Material, Table S6). We considered a correlation between PGS-predicted and measured protein levels as significant at a Bonferroni level of *P*-value <0.05/842. For a PGS variant selection *P*-value cut-off of $p_{\mathrm{PGS}} = 10^{-8}$, the number of protein-PGS that were significantly associated with measured protein levels in both studies was highest (Fig. 3A), while the average number of variants included in any individual PGS increased with r^2_{PGS} (Fig. 3B). We, therefore, chose $p_{\mathrm{PGS}} = 10^{-8}$ and $r^2_{\mathrm{PGS}} = 0.1$ as reference values in the subsequent analyses.

Overall, 268 out of 842 PGS-derived protein levels had a Bonferroni-significant correlation with measured protein levels in both studies (Fig. 3C). For most of these proteins, the correlation between PGS-predicted and measured protein levels was smaller in QBB than in KORA with an average slope of 0.53 for the linear regression between the individual PGS correlations in KORA and QBB. Hence, most PGS performed better in KORA than in QBB.

We then asked how much of the genetic heritability (h^2) was captured by the PGS (r^2). The average protein heritability explained by the PGS was 38.7% in KORA and 35.7% in QBB, estimated as the regression coefficient between r^2 and h^2 (Fig. 4A and B). The explained heritability by the PGS in QBB was 78.6% that of KORA, but with considerable variability depending on the protein in question (Fig. 4C).

A *cis*-variant can regulate protein levels directly via its effect on gene expression, although it can potentially also alter aptamer binding when it is changing the encoded protein's epitope (7). *Cis*-variants often display large effect sizes and can dominate PGS, while *trans*-variants reflect more complex regulatory relationships between proteins and are expected to be involved in more polygenic interactions (31). We, therefore, asked how much of the protein levels was controlled indirectly through *trans*-variants. For this purpose, we excluded all *cis*-variants within a ±10 Mb window of the protein coding gene boundaries and then repeated the process of PGS generation and evaluation as described above. A total of 150 protein PGS for KORA and 152 for QBB retained a significant correlation with the measured protein levels when no *cis*-variants were used, indicating that these protein associations have a significant contribution from *trans*-variants that can be detected by studies that are statistically as highly powered as deCODE (Fig. 4A and B).

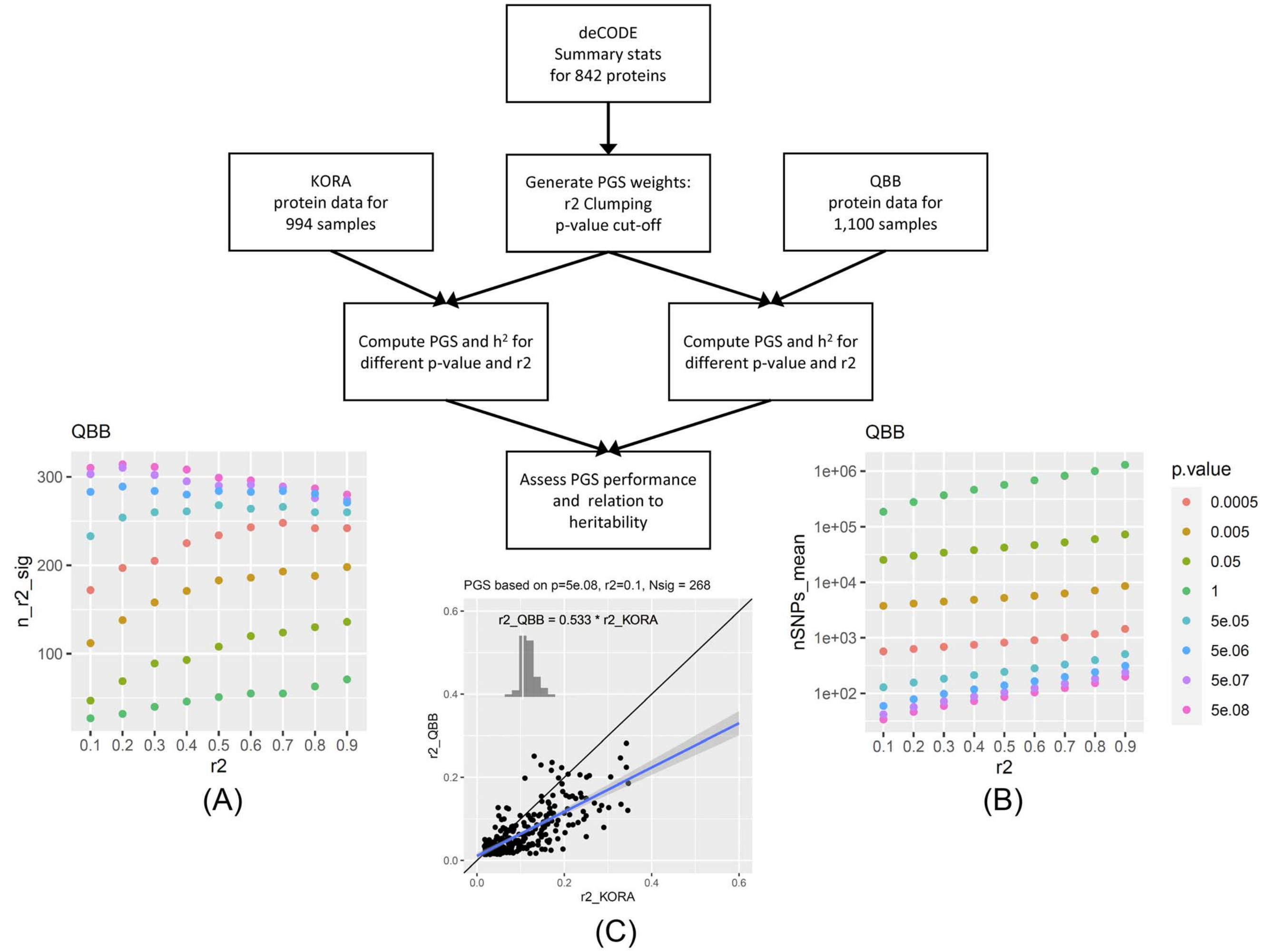

Figure 3. Evaluation of deCODE-derived protein PGS in KORA and QBB. (**A**) A number of Decode-derived PGS (out of 842) evaluated in QBB that correlate with protein levels measured in QBB at a Bonferroni significance level ($P < 0.05/842$), depending on the applied pruning r^2 and significance cut-off (*P*-value); (**B**) the average number of variants included in the PGS as a function of pruning r^2 and cut-off *P*-value; (**C**) correlation of PGS-predicted and measured protein levels in QBB (r^2_QBB) and KORA (r^2_KORA), using a pruning r^2 of 0.1 and a *P*-value cutoff of 5×10^{-8} for the PGS; presented are 268 PGS-derived protein levels that correlate with protein measurements in KORA and QBB at a Bonferroni level of significance; inset: histogram of the frequency distribution of the difference $r^2_{\mathrm{KORA}} - r^2_{\mathrm{QBB}}$.

Discussion

In this investigation, we asked the following questions: do pQTL signals seen in the European KORA and INTERVAL studies replicate in the Arab QBB study? Do the corresponding protein loci display similar genetic structures between the European and the Arab populations? Are there any population specific association signals? And how well do European-derived protein PGS perform in an Arab population?

This is the first large-scale GWAS with proteomics in an Arab population. Availability of WGS data was key for two reasons: one, we were not limited by the current lack of population-specific imputation panels, and two, we could comprehensively identify all potentially causal protein-changing variants, alleviating concerns regarding potential artifacts of affinity proteomics, especially epitope effects when using the SOMAscan platform. Only 10.2% of the *cis*-pQTLs appeared to be attributable to potentially epitope changing variants, almost half of which also have a biological function. We found consistent replication of pQTLs from previous European studies. We attribute the higher replication rate of KORA pQTLs (89.0%) compared to INTERVAL (61.8%) to the fact that the latter study was more highly powered and thus likely to identify more pQTLs with weaker effect sizes that are harder to replicate. For this reason, we did not attempt direct replication of more recent larger pQTL studies, such as deCODE (19) and Pietzner *et al.* (20). The latter was however included in the identification of potentially novel signals through our SNiPA variant annotation, which included these variants.

Few truly novel pQTL signals were found (5.5% of all discovered pQTLs, excluding all previously identified signals at these loci), which is not surprising given the increasing power of the latest pGWAS. For instance, SNP rs5744204 is a pQTL for LBP in our study. This variant has a low MAF (1.3%) in the Caucasian population but is frequent in the Qatar population (15.6%). This pQTL was discovered neither by the KORA and nor by the INTERVAL study but was later reported by the more highly powered study by Pietzner *et al.* (20). This is in agreement with a recent GWAS in the non-European subpopulations of UK Biobank, where analysis of 31 serum and urine biomarker quantitative traits identified 12 novel signals in African and 3 novel signals in South Asian participants, where underlying variants were rare in Europeans with allele frequencies <1% (32).

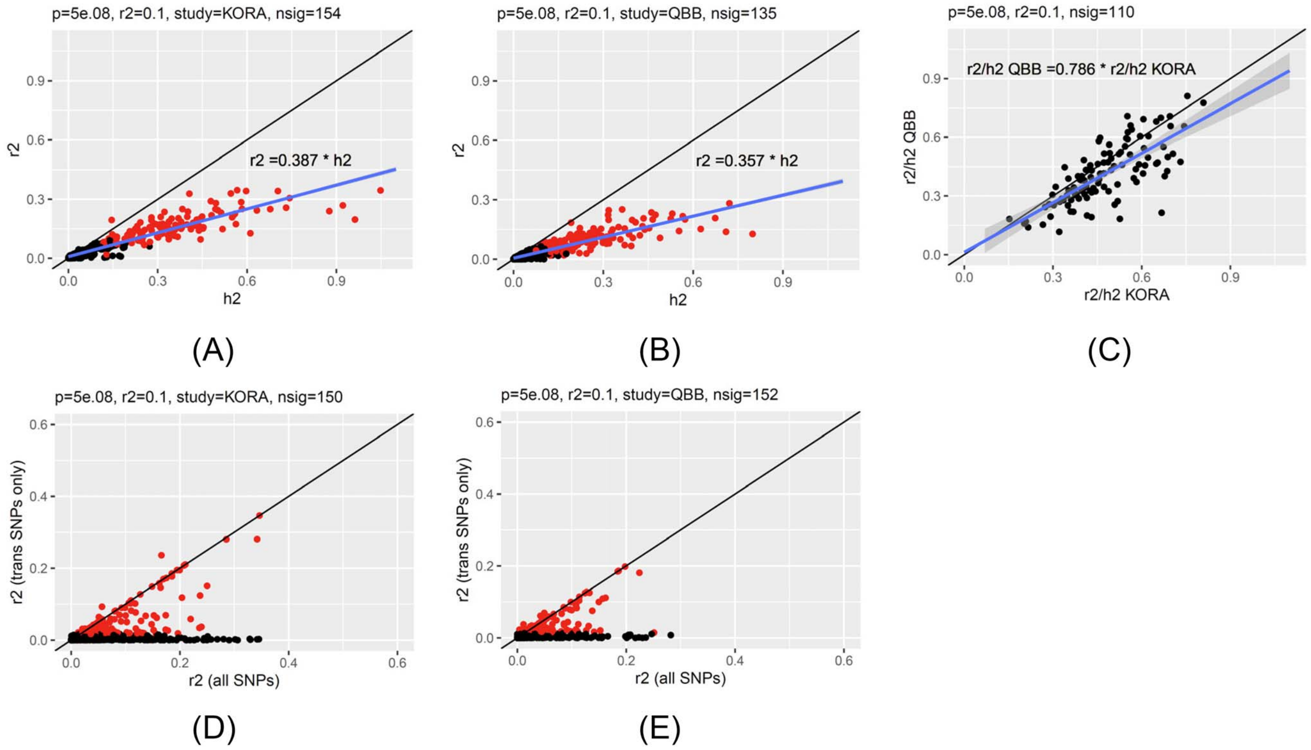

Figure 4. Explained protein heritability between KORA and QBB. Scatterplot of protein variance explained by the PGS (r^2) versus protein heritability (h^2) for KORA (**A**) and QBB (**B**); significant r^2 ($P < 0.05/842$) in red; Scatterplot of the fraction of heritability explained by the PGS (r^2/h^2) in KORA and QBB, only significant proteins shown (**C**); PGS r^2 computed using all variants versus using *trans*-variants only for KORA (**D**) and QBB (**E**), significant r^2 when using *trans*-variants only are in red; note that proteins with solely a *cis*-signal are located near the *x*-axis (black dots) and proteins with a predominant *trans*-signal are located close to the diagonal.

We used the approach of SuSiE-coloc (28) to compare genetic signals between populations at 424 individual genetic regions. While we found at least one shared signal at 81.8% of the loci, it also needs to be acknowledged that many genetic loci host more than one independent pQTL signal. While the approach is designed to work with multiple causal variants, an inspection of the individual plots (Supplementary Material, Fig. S4) revealed cases where improvement is possible. We, therefore, chose not to quantify the number of signals that are not shared between populations, although it is clear that there are many. In summary, we believe it is a fair statement that a majority of the pQTL signals have a shared association signal, but that these are overlain by additional population-specific variants, which contribute to the differences in the performance of the PGS between populations.

We initially expected that protein PGS would capture the polygenic signal of many low effect variants, similar to what was reported by Khera *et al.* (33), who showed that genome-wide PGS emerge as a clinical tool to identify individuals that are at risk for common diseases with risk equivalent to monogenic mutations. However, we found that in our case most PGS were dominated by a few genome-wide significant variants, which is demonstrated by the fact that the strongest correlation between PGS-derived and measured protein levels is obtained when using a *P*-value cutoff of $p_{\mathrm{PGS}} < 10^{-8}$. Adding variants at lower significance levels appears to introduce more noise than information and adding these variants decreases the predictive power of the PGS. We also expected to find several proteins where the PGS is composed of a direct (in-*cis*) and indirect (in-*trans*) contribution, which would correspond to the area located between the *x*-axis and the diagonal in Figure 4D and E. Again, this expectation was not generally met. We hypothesize that the still limited sample size of the study used to generate the PGS is at the root of this observation, indicating that even the size of the deCODE study may still be too small to derive truly polygenic protein PGS. Also, the calculation of protein heritability may have failed for many proteins due to the still limited sample size of current pQTL studies.

Our study has its strengths and limitations. Underrepresentation of Arab populations in imputation panels may bias the results toward European-specific variants. By using WGS data for the GWAS in QBB, we assured that differences were not due to low imputation quality in the Arab population. Furthermore, we only used variants that were polymorphic with a minor allele frequency (MAF) greater than one percent in all three studies, which should further alleviate this problem. As we additionally only retained variants that were present in all three studies in the calculation of the PGS, we may have lost variants that are specific to the European population. This would have weakened the performance of the PGS in the KORA study, implying that differences between the performance of the PGS between the European and Arab populations could be larger if these variants were included.

While relatedness between study participants is not an issue in KORA and INTERVAL, there is a substantial degree of relatedness in QBB. Although we corrected for relatedness in the GWAS part of our study by using dedicated statistical tools, we choose to reduce the number of QBB samples to only unrelated individuals in the comparison of the PGS. While reducing power, this had the positive side-effect that the resulting target cohorts, KORA and QBB, were almost equally powered, which is expected to alleviate possible bias due to differences in cohort sizes.

Previous affinity-proteomics-based pQTL studies were flawed by the problem that an association could potentially be driven by a variant that changed the protein epitope and the pQTL thereby reflected genetic differences of aptamer or antibody binding rather than protein levels (7). Here we quantified the extent of this issue and show that only 10.2% of the pQTLs in our study are potentially affected by protein coding variants. Furthermore, almost half of these protein changing variants were also associated with clinical traits in published GWAS, suggesting that even if they are potentially affecting binding affinity, they also likely have a functional impact. Based on these observations, we estimate that 6.5% of all *cis*-pQTLs discovered using the SOMAscan aptamer technology are affected by pure epitope effects.

Another challenge to aptamer-based affinity proteomics is target specificity. Ideally, each aptamer would be individually validated. This is an ongoing process. The presence of a *cis*-pQTL is considered a strong indicator of target specificity. Williams *et al.* (34) experimentally confirmed target specificity for 93 of the 574 proteins that had at least one pQTL in our study, using pull-down experiments in blood plasma and mass spectrometry for protein identification. 78 (83.9%) out of these confirmed 93 proteins had a *cis*-pQTL in at least one of QBB, KORA or INTERVAL. Moreover, 391 (68.1%) of the 574 proteins identified in our study had a *cis*-pQTL in at least one of the studies. For these proteins, target specificity is thus likely correctly annotated, while further experiments may be required for some of the others, especially for applications where target specificity is essential, such as drug target validation.

Given the fact that we were analyzing WGS data and therefore in principle included all relevant common coding variants, our study suggests that the majority (89.8%) of all pQTLs are driven by non-coding causal variants. This is in line with observations from GWAS with other traits that also found that most causal GWAS variants fall into non-coding regions (35). It is interesting to note that this bias towards non-coding causal variants appears to hold for GWAS with blood circulating protein levels as well.

We applied a readily available methodology to compute PGS. More sophisticated methods may yield slightly different results. Method development is also continuing for the translation of PGS between populations (36). It will be interesting to evaluate these methods in studies with broad molecular phenotypes like ours. Although we cannot publicly share QBB data on an individual phenotype level due to the consent level given by study participants, the data are accessible through direct application to the QBB. In addition, we provide all summary statistics freely through the GWAS catalog to allow future use of our data.

Taken together, our study sheds new light on the genetic architecture of blood circulating pQTLs, which may be relevant for many complex disease associations. It suggests that intermediate molecular traits may be instrumental in estimating the translatability of PGS for clinically relevant traits and disease outcomes and encourages further genetically linked multi-omics studies in underrepresented populations. Such studies, especially if conducted in highly consanguineous populations, like that of Qatar, have the additional potential to identify natural 'knock-outs' that display extreme multi-omics phenotypes by rare variant association analysis (37). Future more highly powered studies with deep molecular phenotypes, such as metabolomics and proteomics, and conducted in a diverse range of populations, are likely to reveal more secrets about the molecular function of the human genome and enable a population-agnostic approach to precision medicine.

Materials and Methods

Study participants

QBB is a population-based study of adult Qatari nationals and long-term residents (living in the country for ≥15 years) (23,24). QBB collects extensive lifestyle, clinical, and biological information on its participants, including metabolomics and proteomics data. The Qatar Genome Project (QGP) builds on data and biosamples from QBB to analyze multiple genetic aspects related to the Qatari population and performs whole genome sequencing (WGS) of Qatari nationals (38). In this study, we analyze 2935 samples from the first batch of QBB that have joint WGS and proteomics data available. All QBB participants signed an informed consent form prior to their participation. The study was approved by Hamad Medical Corporation Ethics Committee and the QBB institutional review board.

Whole genome sequencing

WGS was performed at the Sidra Clinical Genomics Laboratory Sequencing Facility in Doha (Qatar) as previously described (12). Briefly, DNA was extracted from peripheral blood and genomic libraries were sequenced on HiSeq X Ten (Illumina, USA) to achieve a minimum average coverage of 30×, and reads were aligned to the GRCh37 (hs37d53) reference genome using burrows-Wheeler aligner (BWA), and variants were called using GATK 3.4 best practice. The combined variant call format (VCF) file after variant filtering contained 64 997 510 variants for 2935 subjects including 58 713 573 single nucleotide variants (SNVs) and 6 283 936 Indels. Non-autosomal variants (X, Y, mitochondrial DNA), variants with a MAF below 1%, variants with missingness above 10%, variants with Hardy–Weinberg Equilibrium (HWE) $P < 1E\text{-}6$, as well as so-called 'star'-allele variants (variants located in larger indels), were excluded, leaving a total of 10 004 359 variants (SNVs + Indels) for GWAS analysis in 2935 samples. All variant QC filtering steps were performed using Plink v1.90b6.10. Indels were further removed and the dataset was further pruned for $LD > 0.5$ to identify a set of independent markers for estimating principal components (PCs) using Plink and estimating genetic relationship matrix (GRM) using GCTA v1.92.3.

Proteomics data

QBB: blood circulating protein levels were measured using the aptamer-based SOMAscan affinity proteomics platform (Somalogic, Boulder, CO) (39) implemented at Weill Cornell Medicine-Qatar, as previously described (17). Briefly, EDTA-plasma was incubated with bead-coupled epitope-specific aptamers (SOMAmers). Bead-bound proteins were then biotinylated and complexes comprising biotinylated target proteins and fluorescence-labelled SOMAmers were photocleaved and recaptured on streptavidin beads. SOMAmers were then eluted and quantified by hybridization to custom arrays of SOMAmer-complementary oligonucleotides. The resulting raw intensities were processed using different standards as a reference, including hybridization normalization, median signal normalization and signal calibration to control for inter-plate differences. Data for 1305 aptamers were obtained. No samples or data points were excluded. Quality control was performed by repeated measures of two QC samples. The median coefficient of variance (CV) was 0.073 for both QC samples, based on 51 and 54 repeated measures, respectively. 95% of the aptamers had a CV below 0.172 and 0.176, respectively, and 5% had a CV below 0.046 and 0.041, respectively. In other words, half of the assayed proteins had a CV below 0.073 and most (95%) had a CV below 0.176.

KORA: proteomics measurements were performed using the same technology as for QBB at Somalogic but using a previous version of the SOMAscan assay (V3.2) to measure a total of 1129 proteins. Based on 24 measurements each of the two QC samples found that 95% of the aptamers had a CV below 0.136 and 0.104, respectively, and 10% had a CV below 0.033 and 0.027, respectively. The median CV was 0.052 and 0.039, respectively.

Protein annotations

The primary identifiers used in this study are the aptamers ids (SeqId and SomaId in Supplementary Material, Table S7). These identifiers are linked to the proteins that are targeted by the respective aptamers (Uniprot and Entrez gene names). The relation between aptamer identifiers and protein identifiers is not unique: in cases where an aptamer targets a protein complex, multiple protein identifiers are listed (e.g. Complement C1q subcomponent, which is composed of the proteins C1QA, C1QB and C1QC). In other cases, multiple aptamers can target different versions of the same protein, which may lead to duplicated protein identifiers (e.g. APOE). For the definition of *cis*- and *trans*-associations, gene coding regions were obtained for hg build 37. Information on aptamer specificity is from the SOMAscan Assay v4 annotation (version 3.3).

Genome-wide association

The SOMAscan proteomics data (processed RFU values) were log-scaled and residues after regressing against age, gender, the first 10 genetic principal components (PCs), and levels of HSP90 (SeqId 2625-53_4) were computed and then inverse normalized. Four aptamers targeting viral proteins were excluded from the analysis. Mixed linear models were computed using GCTA (26) (v1.92.3 with the—mlma option using GRM as computed above) and summary statistics were saved for further analysis. Median genomic inflation was lambda = 0.993 (range = [0.927, 1.016]). The combined proteome and genome-wide significance level was $P < 3.84\text{E}{-}11$ ($= 5 \times 10^{-8}/1301$). Independent loci were identified and defined as previously described (17) by first clumping all correlated genetic variants on an individual protein trait basis (LD $r^2 > 0.1$, the distance between variants < 10 Mb), keeping always the variant with the strongest association as the sentinel variant, and then grouping all correlated sentinel variants (LD $r^2 > 0.9$) for different protein traits into a single locus. pQTLs for which the sentinel variant was less than 1 Mb distant from or within the boundaries of the protein coding gene were annotated as *cis*-pQTLs, all others as *trans*-pQTLs. In the case of aptamers targeting protein complexes, *cis* annotation was prioritized over *trans*. Regional association plots (RAP) were constructed using LocusZoom software (v 1.4) and with local LD estimates (40).

Locus annotation

The online version of Phenoscanner (41) was used to annotate previously reported pQTLs (http://www.phenoscanner.medschl.cam.ac.uk/, accessed [June 2, 2020]). In one approach all previously published pQTLs were identified that were in LD ($R^2 > 0.8$) with the QBB sentinel variant in at least one of the five populations covered by Phenoscanner, using a relaxed significance level of $P < 10^{-5}$. In a second approach, all pQTLs were identified that were located within a ±10 Mb distance from the QBB sentinel variant at a significance level of $P < 10^{-8}$, regardless of LD. The presence of a pQTL for the same protein trait in at least one of the two approaches was used to annotate that pQTL as 'KNOWN', or as 'NOVEL' otherwise. To update the known pQTLs with the latest associations from the Pietzner *et al.* (20) study, a pre-release version of the SNiPA web server was used (25) (accessed March 2, 2022). Updated overlapping GWAS signals, expression QTLs (eQTLs), metabolomics QTLs (mQTLs) and methylation QTLs (methQTLs) were also annotated and are provided in Supplementary Material, Table S1 and are also accessible online at http://snipa.org.

Variant annotation

Variants of the WGS data from QBB were annotated using Ensembl VEP v99 using the per-gene option. The presence of a potentially protein changing variant for a pQTL was investigated by identifying the most highly correlated variant (LD > 0.8) that changes the protein sequence. Potentially protein changing variants were defined as frame shift variants, in-frame deletion, in-frame insertion, missense variants, splice acceptor variants, splice donor variants, splice region variants, start lost, stop gained or stop loss. To evaluate all possible protein changing variants that could be missed due to standard filtering in GWAS, we used an entire set of 64 997 510 variants, including variants that were filtered in the GWAS for their low quality, in order to identify additional protein changing variants in high LD ($r^2 > 0.8$) with the pQTLs.

Replication of previous pGWAS

Replication of pQTLs reported by two major GWAS using the same technology was attempted. 1980 pQTLs reported by the INTERVAL study were extracted from Supplementary Material, Table S4 of Sun *et al.* (18) and 539 pQTLs identified by the KORA study were extracted from Supplementary Material, Table S1 of Suhre *et al.* (17). In cases where the reported variant was not available in QBB, the most highly correlated tag variant was used. The tag variants were identified for the INTERVAL study using LD from 503 European samples in 1000G phase 3 data (42). Data for 617 and 508 replication attempts were available for INTERVAL and KORA, respectively. The significance level for replication was set to $P < 0.05/617$ and $P < 0.05/508$, respectively.

PGS calculation

Publicly available GWAS summary statistics from the deCODE study (19) were used to derive PGS. The decode summary statistics file were converted from hg38 to hg19 using University of California Santa Cruz (UCSC) liftover tool (43). PGS were computed for KORA using imputed and for QBB using WGS genotype data using PLINK (44). Only variants and proteins that were jointly available in all three studies were included. PGS variants were selected by clumping correlated variants based on LD (r^2_{PGS}) and discarding all associations below a fixed *P*-value cut-off (p_{PGS}). The LD values are computed using 503 European samples in 1000G phase 3 data. PGS performance was quantified using Pearson correlations between PGS-derived and observed protein levels in KORA and QBB using individual level SOMAscan data from both studies. To determine the sensitivity of the PGS performance, we repeated the PGS generation process using a range of significance cutoffs ($p_{\text{PGS}} = 1, 0.1, 0.01, \ldots, 10^{-8}$) and clumping r^2 values ($r^2_{\text{PGS}} = 0.1, 0.2, \ldots, 1.0$).

Supplementary Material

Supplementary Material is available at *HMGJ* online.

Acknowledgements

QBB and Qatar Genome Program are both Research, Development & Innovation entities within Qatar Foundation for Education, Science and Community Development. We gratefully acknowledge the contribution of all members of field staff conducting the KORA and the QBB studies, and in particular thank all study participants for donating their time and blood.

Conflict of Interest statement. None declared.

Ethics Statements

QBB: all participants signed an informed consent form prior to their participation, and the study was approved by Hamad Medical Corporation Ethics Committee and QBB institutional review board. Data were accessed under project numbers QBB-IRB_E -2018-QGP-PUB-009 and E-2017-QF-QGP -RES-PUB-009-0015. KORA: all study participants have given written informed consent and the study was approved by the Ethics Committee of the Bavarian Medical Association. INTERVAL: only publicly available summary statistics data were used; see ethics statement in Sun *et al.* (18). deCODE: only publicly available summary statistics data were used; see ethics statement in Ferkingstad *et al.* (19).

Data Availability

Full summary statistics from this study have been deposited in the NHGRI-EBI Catalog of human GWAS and can be accessed through https://www.ebi.ac.uk/gwas/ under accession code [link to be inserted, data are presently available without password protection at https://wcmq.box.com/s/vkhhqz4z1ijdew5p2xfvi3ta0rww5n30]. The informed consent given by the QBB and KORA study participants does not cover the posting of participant level phenotype and genotype data in public databases. Access to participant level QBB/QGP genotype and phenotype data can be obtained through an established ISO-certified process by submitting a project request at https://www.qatarbiobank.org.qa/research/how-to-apply which is subject to approval by the QBB IRB committee. Data for KORA are available upon request from KORA (https://helmholtz-muenchen.managed-otrs.com/external). Requests are submitted online and are subject to approval by the KORA board. KORA and INTERVAL summary statistics can be freely accessed at https://www.ebi.ac.uk/gwas/downloads/summary-statistics and those of deCODE at https://www.decode.com/summarydata/.

Funding

Biomedical Research Program at Weill Cornell Medicine in Qatar, a program funded by the Qatar Foundation. O.M.E.A. and K.S. are supported by Qatar National Research Fund (QNRF, grant NPRP11C-0115-180010). O.M.E.A. is also supported by a QNRF (grant PPM 03-0324-19003). A.B. is supported by NPRP-Standard (NPRP-S) Twelfth (12th) Cycle grant NPRP12S-0227-190173 from the Qatar National Research Fund (a member of Qatar Foundation). Qatar Biobank and Qatar Genome Program are supported by Qatar Foundation. M.A. received funding through Duke University/Helmholtz Zentrum München from the National Institutes of Health/National Institute on Aging through grants RF1AG058942, RF1AG059093, U01AG061359, U19AG063744 and R01AG069901. The statements made herein are solely the responsibility of the authors. 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 and by the State of Bavaria. This project is also funded by the Bavarian State Ministry of Health and Care through the research project DigiMed Bayern (www.digimed-bayern.de).

Authors' Contributions

K.S. designed and supervised the study; G.T. performed the genome-wide association study and conducted the primary data analysis; M.A., A.B., O.M.E.A. contributed to data analysis and interpretation; C.G., J.G., F.S., H.G., A.P. provided analyses tools and material. K.S. wrote the manuscript. Members of The Qatar Genome Program Research Consortium contributed to subject recruitment, phenotyping, data acquisition and WGS. All authors approved the final version of the manuscript.

References

1. MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J. *et al.* (2017) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.*, **45**, D896–D901.
2. Davey Smith, G. and Hemani, G. (2014) Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.*, **23**, R89–R98.
3. Torkamani, A., Wineinger, N.E. and Topol, E.J. (2018) The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.*, **19**, 581–590.
4. Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.-R., Duncan, L., Perry, J.R.B., Patterson, N., Robinson, E.B. *et al.* (2015) An atlas of genetic correlations across human diseases and traits. *Nat. Genet.*, **47**, 1236–1241.
5. Kronenberg, F. (2012) In Suhre, K. (ed), *Genetics Meets Metabolomics: From Experiment to Systems Biology*. Springer New York, New York, NY, pp. 255–264.
6. GTEx-Consortium (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.
7. Suhre, K., McCarthy, M.I. and Schwenk, J.M. (2020) Genetics meets proteomics: perspectives for large population-based studies. *Nat. Rev. Genet.*, **22**, 19–37.
8. Suhre, K. and Gieger, C. (2012) Genetic variation in metabolic phenotypes: study designs and applications. *Nat. Rev. Genet.*, **13**, 759–769.
9. Plenge, R.M., Scolnick, E.M. and Altshuler, D. (2013) Validating therapeutic targets through human genetics. *Nat. Rev. Drug Discov.*, **12**, 581–594.
10. Gurdasani, D., Barroso, I., Zeggini, E. and Sandhu, M.S. (2019) Genomics of disease risk in globally diverse populations. *Nat. Rev. Genet.*, **20**, 520–535.
11. Lambert, S.A., Gil, L., Jupp, S., Ritchie, S.C., Xu, Y., Buniello, A., McMahon, A., Abraham, G., Chapman, M., Parkinson, H. *et al.* (2021) The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. *Nat. Genet.*, **53**, 420–425.
12. Thareja, G., Al-Sarraj, Y., Belkadi, A., Almotawa, M., Suhre, K. and Albagha, O.M.E. (2021) Whole genome sequencing in the Middle Eastern Qatari population identifies genetic associations with 45 clinically relevant traits. *Nat. Commun.*, **12**, 1250.
13. Sinnott-Armstrong, N., Tanigawa, Y., Amar, D., Mars, N., Benner, C., Aguirre, M., Venkataraman, G.R., Wainberg, M., Ollila, H.M., Kiiskinen, T. *et al.* (2021) Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat. Genet.*, **53**, 185–194.
14. Solomon, T., Lapek, J.D., Jr., Jensen, S.B., Greenwald, W.W., Hindberg, K., Matsui, H., Latysheva, N., Braekken, S.K., Gonzalez, D.J.,

Frazer, K.A., Smith, E.N. and Hansen, J.B. (2018) Identification of common and rare genetic variation associated with plasma protein levels using whole-exome sequencing and mass spectrometry. *Circulation. Genomic Precision Med.*, **11**, e002170.

15. Emilsson, V., Ilkov, M., Lamb, J.R., Finkel, N., Gudmundsson, E.F., Pitts, R., Hoover, H., Gudmundsdottir, V., Horman, S.R., Aspelund, T. *et al.* (2018) Co-regulatory networks of human serum proteins link genetics to disease. *Science*, **361**, 769–773.
16. Folkersen, L., Gustafsson, S., Wang, Q., Hansen, D.H., Hedman, Å.K., Schork, A., Page, K., Zhernakova, D.V., Wu, Y., Peters, J. *et al.* (2020) Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nat. Metab.*, **2**, 1135–1148.
17. Suhre, K., Arnold, M., Bhagwat, A.M., Cotton, R.J., Engelke, R., Raffler, J., Sarwath, H., Thareja, G., Wahl, A., DeLisle, R.K. *et al.* (2017) Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat. Commun.*, **8**, 14357.
18. Sun, B.B., Maranville, J.C., Peters, J.E., Stacey, D., Staley, J.R., Blackshaw, J., Burgess, S., Jiang, T., Paige, E., Surendran, P. *et al.* (2018) Genomic atlas of the human plasma proteome. *Nature*, **558**, 73–79.
19. Ferkingstad, E., Sulem, P., Atlason, B.A., Sveinbjornsson, G., Magnusson, M.I., Styrmisdottir, E.L., Gunnarsdottir, K., Helgason, A., Oddsson, A., Halldorsson, B.V. *et al.* (2021) Large-scale integration of the plasma proteome with genetics and disease. *Nat. Genet.*, **53**, 1712–1721.
20. Pietzner, M., Wheeler, E., Carrasco-Zanini, J., Cortes, A., Koprulu, M., Wörheide, M.A., Oerton, E., Cook, J., Stewart, I.D., Kerrison, N.D. *et al.* (2021) Mapping the proteo-genomic convergence of human diseases. *Science*, **374**, eabj1541.
21. Ritchie, S.C., Liu, Y., Lambert, S.A., Teo, S.M., Scepanovic, P., Marten, J., Zahid, S., Chaffin, M., Abraham, G., Ouwehand, W.H. *et al.* (2021) Integrative analysis of the plasma proteome and polygenic risk of cardiometabolic diseases. *Nature Metab.*, **3**, 1476–1483.
22. Zaghlool, S.B., Sharma, S., Molnar, M., Matías-García, P.R., Elhadad, M.A., Waldenberger, M., Peters, A., Rathmann, W., Graumann, J., Gieger, C., Grallert, H. and Suhre, K. (2021) Revealing the role of the human blood plasma proteome in obesity using genetic drivers. *Nat. Commun.*, **12**, 1279.
23. Al Kuwari, H., Al Thani, A., Al Marri, A., Al Kaabi, A., Abderrahim, H., Afifi, N., Qafoud, F., Chan, Q., Tzoulaki, I., Downey, P. *et al.* (2015) The Qatar Biobank: background and methods. *BMC Public Health*, **15**, 1208.
24. Al Thani, A., Fthenou, E., Paparrodopoulos, S., Al Marri, A., Shi, Z., Qafoud, F. and Afifi, N. (2019) Qatar biobank cohort study: Study design and first results. *Am. J. Epidemiol.*, **188**, 1420–1433.
25. Arnold, M., Raffler, J., Pfeufer, A., Suhre, K. and Kastenmuller, G. (2015) SNiPA: an interactive, genetic variant-centered annotation browser. *Bioinformatics*, **31**, 1334–1336.
26. Yang, J., Lee, S.H., Goddard, M.E. and Visscher, P.M. (2011) GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.*, **88**, 76–82.
27. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C. and Plagnol, V. (2014) Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.*, **10**, e1004383.
28. Wallace, C. (2021) A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLoS Genet.*, **17**, e1009440.
29. Kanai, M., Ulirsch, J.C., Karjalainen, J., Kurki, M., Karczewski, K.J., Fauman, E., Wang, Q.S., Jacobs, H., Aguet, F., Ardlie, K.G. *et al.* (2021) Insights from complex trait fine-mapping across diverse populations. medRxiv. https://doi.org/2021.2009.2003.21262975.
30. Zhang, J., Dutta, D., Köttgen, A., Tin, A., Schlosser, P., Grams, M.E., Harvey, B., Consortium, C., Yu, B., Boerwinkle, E. et al (2022) Plasma proteome analyses in individuals of European and African ancestry identify *cis*−pQTLs and models for proteome-wide association studies. *Nat Genet.*, **54**, 593–602.
31. Boyle, E.A., Li, Y.I. and Pritchard, J.K. (2017) An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*, **169**, 1177–1186.
32. Sun, Q., Graff, M., Rowland, B., Wen, J., Huang, L., Lee, M.P., Avery, C.L., Franceschini, N., North, K.E., Li, Y. *et al.* (2022) Analyses of biomarker traits in diverse UK biobank participants identify associations missed by European-centric analysis strategies. *J. Hum. Genet.*, **67**, 87–93.
33. Khera, A.V., Chaffin, M., Aragam, K.G., Haas, M.E., Roselli, C., Choi, S.H., Natarajan, P., Lander, E.S., Lubitz, S.A., Ellinor, P.T. and Kathiresan, S. (2018) Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.*, **50**, 1219–1224.
34. Williams, S.A., Kivimaki, M., Langenberg, C., Hingorani, A.D., Casas, J.P., Bouchard, C., Jonasson, C., Sarzynski, M.A., Shipley, M.J., Alexander, L. *et al.* (2019) Plasma protein patterns as comprehensive indicators of health. *Nat. Med.*, **25**, 1851–1857.
35. Zhang, F. and Lupski, J.R. (2015) Non-coding genetic variants in human disease. *Hum. Mol. Genet.*, **24**, R102–R110.
36. Weissbrod, O., Kanai, M., Shi, H., Gazal, S., Peyrot, W.J., Khera, A.V., Okada, Y., Matsuda, K., Yamanashi, Y., Furukawa, Y. *et al.* (2022) Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. *Nat. Genet.*, **54**, 450–458.
37. Narasimhan, V.M., Xue, Y. and Tyler-Smith, C. (2016) Human knockout carriers: dead, diseased, healthy, or improved? *Trends Mol. Med.*, **22**, 341–351.
38. Mbarek, H., Devadoss Gandhi, G., Selvaraj, S., Al-Muftah, W., Badji, R., Al-Sarraj, Y., Saad, C., Darwish, D., Alvi, M., Fadl, T. *et al.* (2022) Qatar genome: insights on genomics from the Middle East. *Hum. Mutat.*, **43**, 499–510.
39. Gold, L., Ayers, D., Bertino, J., Bock, C., Bock, A., Brody, E.N., Carter, J., Dalby, A.B., Eaton, B.E., Fitzwater, T. *et al.* (2010) Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoS One*, **5**, e15004.
40. Pruim, R.J., Welch, R.P., Sanna, S., Teslovich, T.M., Chines, P.S., Gliedt, T.P., Boehnke, M., Abecasis, G.R. and Willer, C.J. (2010) LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics*, **26**, 2336–2337.
41. Staley, J.R., Blackshaw, J., Kamat, M.A., Ellis, S., Surendran, P., Sun, B.B., Paul, D.S., Freitag, D., Burgess, S., Danesh, J., Young, R. and Butterworth, A.S. (2016) PhenoScanner: a database of human genotype-phenotype associations. *Bioinformatics*, **32**, 3207–3209.
42. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A. and Abecasis, G.R. (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
43. Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F. *et al.* (2006) The UCSC genome browser database: update 2006. *Nucleic Acids Res.*, **34**, D590–D598.
44. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. and Sham, P.C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.