A genome-wide association study of mass spectrometry proteomics using the Seer Proteograph platform

Karsten Suhre[1,2,], Qingwen Chen[3], Anna Halama[1,2], Kevin Mendez[3], Amber Dahlin[3], Nisha Stephan[1], Gaurav Thareja[1], Hina Sarwath[4], Harendra Guturu[5], Varun B. Dwaraka[6], Serafim Batzoglou[5], Frank Schmidt[4], Jessica A. Lasky-Su[3,*]*

[1] Bioinformatics Core, Weill Cornell Medicine-Qatar, Education City, 24144 Doha, Qatar

[2] Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY 10021, USA.

[3] Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, U.S.A.

[4] Proteomics Core, Weill Cornell Medicine-Qatar, Education City, 24144 Doha, Qatar

[5] Seer, Inc., Redwood City, Redwood City, CA 94065, U.S.A.

[6] TruDiagnostic, Inc., Lexington, KY, U.S.A.

[*] Correspondence to K.S. (kas2049@qatar-med.cornell.edu) and J.A.L. (rejas@channing.harvard.edu)

ABSTRACT

Genome-wide association studies (GWAS) with proteomics are essential tools for drug discovery. To date, most studies have used affinity proteomics platforms, which have limited discovery to protein panels covered by the available affinity binders. Furthermore, it is not clear to which extent protein epitope changing variants interfere with the detection of protein quantitative trait loci (pQTLs). Mass spectrometry-based (MS) proteomics can overcome some of these limitations. Here we report a GWAS using the MS-based Seer Proteograph^{TM} platform with blood samples from a discovery cohort of 1,260 American participants and a replication in 325 individuals from Asia, with diverse ethnic backgrounds. We analysed 1,980 proteins quantified in at least 80% of the samples, out of 5,753 proteins quantified across the discovery cohort. We identified 252 and replicated 90 pQTLs, where 30 of the replicated pQTLs have not been reported before. We further investigated 200 of the strongest associated cis-pQTLs previously identified using the SOMAscan and the Olink platforms and found that up to one third of the affinity proteomics pQTLs may be affected by epitope effects, while another third were confirmed by MS proteomics to be consistent with the hypothesis that genetic variants induce changes in protein expression. The present study demonstrates the complementarity of the different proteomics approaches and reports pQTLs not accessible to affinity proteomics, suggesting that many more pQTLs remain to be discovered using MS-based platforms.

INTRODUCTION

Genome-wide association studies (GWAS) with proteomics leverage the naturally occurring genetic variation in human populations and link differences between individual genomes to their effects on the proteome and beyond [1]. Protein quantitative trait loci (pQTLs) are central in the drug discovery process as they provide supporting evidence for drug target identification and hypothesis generation on their modes of action [2]. Most large-scale GWAS so far have been conducted using affinity proteomics platforms [3-12]. The largest studies to date are from deCODE using the SOMAscan platform with 4,907 aptamers in 35,559 Icelandic samples from Icelanders [13] and from the UKB-PPP consortium using the Olink platform with dual antibodies targeting 2,923 proteins in 54,219 samples from participants of the UK Biobank [14]. These studies reported thousands of pQTLs that are now available for further exploitation, such as Mendelian randomization (MR) experiments, to identify new drug targets and to further their development.

However, it should be noted that SOMAscan and Olink pQTLs represent genetic associations with protein binding affinity, rather than protein expression. This is because protein altering variants (PAVs) can modify the affinity binding epitopes of the proteins, thereby leading to genotype-dependent read-outs that do not correspond to real changes in protein levels [15]. Such epitope effects can invalidate conclusions drawn from MR experiments, as their basic hypothesis requires that changes in the exposure, i.e. the protein expression levels, are causal for changes in the disease outcome. Epitope effects can also skew the prediction of protein levels using polygenic scores and confound correlations with other omics modalities. Therefore, it is important to validate key pQTLs on an independent platform that is immune against epitope effects, such as MS-based proteomics.

A few smaller-scale MS-based GWAS have been previously reported. Johansson et al. [16] identified and quantified the abundance of 1,056 tryptic-digested peptides, representing 163 proteins in the plasma of 1,060 individuals from two population-based cohorts. Xu et al. [17] conducted a GWAS for 304 proteins measured by SWATH-MS in blood serum from 2,958 Han Chinese individuals. Niu et al. [18] performed a GWAS for 420 proteins using MS-based proteomics in blood plasma from 1,914 children and adolescents with a replication in 558 adults. However, these studies were all limited to a small number of proteins.

Here we conduct a GWAS using the MS-based Proteograph™ platform (*Seer, Inc.*) for deep, unbiased proteomics [15,19,20]. The Seer technology increases proteome coverage by using nanoparticle enrichment, followed by a data-independent acquisition protocol implemented on a Bruker timsTOF Pro 2 mass spectrometer (*Bruker Daltonics*). We have previously shown that the protein readouts of the Proteograph platform can reliably distinguish between epitope- and protein expression QTLs when a specific data analysis protocol is applied that eliminates PAV-containing peptides. We extended this approach here to a larger study cohort and a full GWAS [15]. We analysed 1,980 proteins that were quantified in at least 80% of the samples, out of 5,753 proteins quantified across a discovery cohort of 1,260 participants of a diverse American ethnic background and a replication phase using 325 samples from participants of mainly Arab, Indian, and Filipino ethnic backgrounds. We then validated the protein associations with age and sex against those reported by the two largest affinity proteomics GWAS. Finally, we investigated the lead cis-pQTLs reported by the SOMAscan and Olink platforms for potential epitope effects.

RESULTS

A GWAS with the Seer Proteograph platform identified 252 and replicated 90 pQTLs.

We conducted a GWAS with 1,980 proteins detected in >80% of the 1,260 samples from the Tarkin study (**Table S1**, see Methods). A total of 252 independent protein associations reached a Bonferroni level of significance ($p < 2.5x10^{-11}$), involving 224 genetic loci for 152 different proteins (**Figure 1** and **Table S2**). Replication was attempted using 325 samples from the QMDiab study with matching genotype and proteomics data. A pQTL was considered replicated if its association reached a significance level of p < 0.05 / 252 and had concordant effect direction. A total of 90 pQTLs satisfied these criteria. Of the 94 pQTLs calculated to have 80% replication power, determined by sampling, a total of 60 pQTLs (63.8%) replicated, and most of the non-replicated pQTLs also had concordant directionality (**Figure 2A**). Differences in allele frequencies of the pQTLs between Tarkin and QMDiab were generally in the range of +/-10% allele frequency (**Figure 2B**).

MS-based proteomics platforms generate a rich set of readouts, including quantifications of multiple peptides issued from a same protein, often repeatedly measured at multiple precursor charges. We generated visualizations of this data stratified by genotype for each of the 252 pQTLs (**Figure S1**). An example is given in **Figure 3**, where the associations with the four most frequently quantified peptides per protein are shown together with the derived protein level associations, both for Tarkin and QMDiab. Additionally, we present plots of PAV containing peptides, in cases they were detected for the protein in question. These plots provide individual verifiable experimental evidence that further supports the validity of these pQTLs.

We then asked which MS-based pQTLs had been identified before, and which were novel. Summary statistics from the deCODE SOMAscan study [13] and the UKB-PPP Olink study [14] were used to identify overlapping associations with previously reported pQTLs. Of the 252 pQTLs identified in this study, 65 were found by both platforms, 43 were reported by SOMAscan alone and 33 were seen by Olink alone. Four pQTLs were not reported by either of these studies, but were identified using PhenoScanner [21] as a pQTL in some other study. This leaves 107 pQTLs that have not been reported before, including 30 of the 90 replicated pQTLs (**Table 1**), suggesting an expected novel discovery rate when using the Proteograph platform of about one in three compared to existing affinity pQTLs.

Age and sex associations were concordant between the affinity and MS based proteomics platforms.

We computed the associations between all 1,980 protein readouts with age, sex, and the ten genotype principal components (**Table S3**). We then asked whether the associations with age and sex were concordant between the two affinity platforms and between the affinity and the MS proteomics platforms. A total of 507 proteins were quantified on all three platforms (**Figure S1A and Table S4**). The overall effect sizes were concordant for most of the proteins between all three platforms (**Figure 4**), although with some exceptions for the associations with age. Overall, 62 proteins shared significant associations with sex across all three platforms (**Figure S1B**), while 110 proteins exhibited significant associations with age (**Figure S1C**).

One third of affinity proteomics pQTLs are potentially affected by epitope effects.

The deCODE SOMAscan study [13] and the UKB-PPP Olink study [14] reported 1,881 and 1,917 cis-pQTLs, respectively (**Table S5 & S6**). For 1,041 and 1,415 of these pQTLs matching genetic variants were available in Tarkin and QMDiab. Out of these, 322 and 374 pQTLs also had the matching protein quantified in Tarkin with a missing rate of less than 20%. For these pQTLs the genetic associations with the matching protein and all corresponding peptide levels were computed using the Proteograph data, both for the Tarkin and the QMDiab study (**Table S7 & S8**). Note that most pQTLs without a matching variant in Tarkin had minor allele frequencies below 5% and were excluded from our analysis.

Quantification of protein levels from peptide intensities is sometimes a limiting factor in MS proteomics and prone to large uncertainties. We therefore integrated information from individual peptide measurements. For each pQTL the association data was integrated into what we refer to here as the MS peptide association (MSPA) score. This score is designed as a proxy for the likelihood of a protein expression signal being observed in the MS proteomics data and combines association signals from all peptides observed for a given protein. Note that PAV containing peptides had been removed from our MS peptide library. We defined the MSPA score as follows: All peptides were attributed a weight using the number of individual peptide detections for that protein, divided by the sum of all peptide detections for that protein. Repeated peptide detections with different precursor charges were excluded, retaining the peptide with the highest number of detections. The weights of all peptide associations where the 99% confidence interval of the beta estimate did not contain the zero were then summed up. An MSPA score of one therefore represents a case where all peptides support a genetic association with the protein expression level, and an MSPA score of zero indicates cases where the data provides no statistical support for such an association, suggesting either a lack of statistical power to detect an association or the presence of an epitope effect. As association directionality was not included into the score, and to identify possible artifacts, the associations for all analysed pQTLs were visualized for additional manual inspections; no inconsistencies were found (**Figure S2&S3**).

The MSPA score of each pQTL was then plotted against the rank of the p-value of the original pQTL from the affinity proteomics study. As shown in **Figure 5**, a sigmoid distribution running from the upper left to the lower right corner of the plot may be discerned for both studies. This curve approximately represents the MSPA score expected for a protein expression QTL as a function of pQTL rank as statistical power decreases with rank. Roughly, the first 100 pQTLs of each study appear to have sufficient power to replicate in Tarkin. Within these 100 strongest *cis*-pQTLs, both studies have almost the same proportions of protein expression QTLs (35% SOMA, 37% Olink) and the same number of likely epitope effect driven pQTLs (33% for SOMA and Olink). Hence, these plots suggest that about one third of the affinity proteomics pQTLs are possibly due to epitope effects, one third are reproducible using MS proteomics in a study of our current size, and one third fall into a "grey zone" where power may be an issue, but hybrids of epitope- and expression-pQTLs are also possible, that is, cases where a genetic variant interferes with the affinity binding, but at the same time affects protein expression via some biological feedback mechanism.

To further support the validity of the MSPA score as a proxy for the detection/non-detection of a genetic association and its potential to identify true positive protein expression pQTLs, we selected all pQTLs that were reported on the same genetic variant by deCODE and UKBPPP, and for which also matching protein and genetic data were also available in Tarkin and QMDiab. We used this set of 46 pQTLs as a gold standard for comparing effect sizes and directionality without the caveat of having to use proxy SNPs (**Table S9**). A few things were notable when looking at this set of pQTLs. Firstly, in contrast to the overall distribution of the MSPA scores in **Figures 5A and 5B**, where there were many high ranking pQTLs with low and zero MSPA scores present, the top ranking pQTLs in this set all had high MSPA scores (**Figures 5C and 5D)**, suggesting that none of these pQTLs were affected by epitope effects. This inference is further supported by the near perfect correlation of the effect sizes between the Olink and the SOMAscan platforms (**Figure 5E**). Considering that these pQTLs had been pre-selected based on the requirement that they were detected on both affinity platforms, it appears that a pQTL being detected by both affinity platforms is also a strong indicator for a true protein expression QTL and the absence of an epitope effect. This is a reasonable assumption because there is low likelihood that two different affinity binders target the same surface area of a protein and lead to a similar epitope readout.

If these 46 pQTLs are mostly devoid of epitope effects, the dependence of the MSPA values on the association strength (rank) should follow the power of each pQTL to be detected by the Tarkin study, supporting our earlier assumption that the high-ranking pQTLs with low MSPA scores in **Figure 5A** and **5B** are epitope-effect driven pQTLs. We therefore asked whether a candidate epitope-changing variant was identifiable in all these cases (the pQTLs in the red zone of **Figure 5A** and **5B**). We manually queried the Ensembl database [22] for coding variants that were potentially epitope-changing (**Table S10**) and found that for 23 out of 33 SOMA and 25 out of 33 Olink pQTLs such a variant had been reported (requiring LD r^2>0.8). The lead pQTL SNP or a SNP in perfect LD of r^2=1 was apparently epitope-changing in all but four of the 48 cases. Additionally, in 11 out of these 48 cases, a PAV-containing peptide was also detected on the Proteograph platform, with heterozygotes having about half the protein level, showing that the absence of a protein expression pQTL is not due to limitations in the quantification of the peptides, and confirming the presence of the protein variant in blood. In five cases (CPN2, SERPINA1, ENO3, HDGF, APOBR) both alleles, reference (REF) and alternate (ALT), were detected and showed significant associations with the coding variant, while all other peptides did not associate with the variant (for an example see **Figure 7**).

DISCUSSION

This study presents the first comprehensive GWAS using the MS-based Seer Proteograph platform, accompanied by a full replication, and employing a proteomics data analysis protocol that accounts for genetic variants within the analysed peptide [15]. Our methodology not only identified new pQTLs on proteins previously unassessed by affinity proteomics platforms, but also examined previously reported affinity pQTLs for confounding due to epitope-altering variants. We estimate that about one-third of the 200 pQTLs evaluated were influenced by such effects. However, due to limitations in statistical power that restricted evaluation to only the strongest affinity pQTLs, which are the ones most likely to be enriched for epitope effects due to ascertainment bias, this estimate should be considered as an upper bound.

We reported a total of 252 pQTLs, with 90 successfully replicated in an independent population. While the replication rate of 63.8% for sufficiently powered pQTLs is high, it falls below the theoretically expected 80%. We attribute this difference to genetic and probably also lifestyle differences between the discovery and replication cohorts, which are ethnically very distinct. Nonetheless, this diversity also enhances the robustness and translatability of the replicated pQTLs across populations.

The use of a proteome library that accounts for PAVs was central to our study. Without using this approach, a very high number of false positive pQTLs would have been detected, as we previously discussed [15]. Traditional proteomics data analysis methods often rely on a limited peptide library for protein quantification, and the presence of a single peptide with a large effect can skew the quantification. A PAV containing peptide would not be detected in homozygotes of the alternate allele and heterozygotes would have half the peptide level. The inclusion of PAV containing peptides into the protein quantification would hence lead to the equivalent of an epitope effect, that is, a pQTL signal where in reality, there is no genotype dependence of the protein expression level. Also, PAV peptides corresponding to the alternate allele are not present in an in-silico digest library of the standard UniProt database. Indeed, using data from a standard proteomics data processing run, we observed cases where fragment spectra of these peptides miss-matched to peptides from other proteins and in extreme cases led to false protein identifications.

We used the associations with sex and age to externally validate the comparability of the affinity and MS proteomics readouts to predict non-genetic outcomes. Although our MS-based study was less powered compared to the two much larger affinity proteomics studies, the scatterplots of the effect sizes between the two large studies and one of the large studies and our study are largely comparable. Key signals, such as the association of sex hormone binding globulin (SHBG) with sex and leptin (LEP) with age, were significant on all platforms. There are more age-related associations with conflicting effect sizes. These may be population specific and lifestyle related effects. Noteworthy is also the higher number of positive associations with age that is present across all three platforms, which we can only speculate about.

Our study also has some caveats. The interpretation of absence of pQTL signal in MS proteomics as indicative of epitope effects necessitates formal power analysis. However, such an approach is challenging because it requires precisely estimating the effect sizes for MS-based data from associations using affinity proteomics, whereas effect sizes are measured in relative units of standard deviations within the study population that do not translate between platforms and studies. Thus, we limited our analysis to the 100 strongest pQTLs in each study, under the assumption that these are sufficiently powered to replicate using MS proteomics, if the signal is truly based on protein expression rather than affinity binding. We support this argument using the distribution of the MSPA scores as discussed around **Figure 5**.

As we interpret the absence of a pQTL signal using an MS platform as an indicator of a potential epitope effect, we applied a stringent missingness criterion (<20%). The total number of 1,980 proteins analysed in this GWAS are therefore lower than the 5,753 proteins that were quantified in any sample across the study.

Additionally, the matching of proteins between platforms using UniProt identifiers can be challenging, as affinity proteomics platforms sometimes report multiple protein identifiers when they are targeting protein complexes or use the UniProt identifier specific to the proteoform employed to generate the affinity binder while alternate proteoforms may be included in the MS library, occasionally leading to annotations by protein groups rather than specific proteins. Mapping of UniProt ids to gene names can also be challenging, especially when the versions of the underlying databases do not match between studies.

While we reported many novel pQTLs, about one third of the identified pQTLs, we opted not to highlight any new biologically relevant findings. We believe that the biomedical relevance of pQTLs has already been amply shown in many of the previous pQTL studies, and that cherry-picking one or two new highlights would distract from the central message of our paper, which is to demonstrate the complementarity of MS proteomics to affinity approaches in (1) validating associations that may be driven by epitope effects and (2) substantially extending the panel of proteins accessible to pQTL studies.

METHODS

Ethics. The original Tarkin study was reviewed and approved by the IRCM IRB and the Mass General Brigham IRB. Participants provided written informed consent to take part in the study. The WCMQ IRB determined that use of the Tarkin data for the present project does not meet the definition of human research for this study (IRB document HRP-532). The QMDiab study was approved by the institutional research boards of Weill Cornell Medicine – Qatar under protocol #2011-0012 and of Hamad Medical Corporation under protocol #11131/11 and complies with all relevant ethical regulations. For forthgoing work with the study a non-human subject research determination was obtained. The study design and conduct complied with all relevant regulations regarding the use of human study participants and was conducted in accordance with the criteria set by the Declaration of Helsinki.

Cohorts. The samples used in the Tarkin study were obtained from the Massachusetts General Brigham (MGB) Biobank. Joint phenotype and genotype data were available for 1,260 samples, comprised of 662 females and 598 males with an age range of 23-99 years (median 70 years, mean 67.2 years). 1,057 of the participants were white. This subset of MGB samples together with its deep omics characterization is referred to here as the “Tarkin study” in this paper [23]. For replication, a total of 345 previously unthawed citrate blood plasma samples from participants of the Qatar Metabolomics study of Diabetes (QMDiab), including adult female and male participants of predominantly Arab, Indian, and Filipino ethnic background, with and without diabetes in an age range from 18-80 were assayed using the Proteograph platform (*Seer Inc.*) [15,24,25].

Genotyping. Imputed genotype data for 1,980 samples of the Tarkin study was received on a per-chromosome basis in vcf format (build 37, imputed using Minimac3, no indels). The genotype data was filtered for biallelic variants and variant names were standardized using bcftools (version 1.16), converted to plink format and filtered using plink2 (version v2.00a5LM) with the options --geno 0.1 --mac 10 --maf 0.05 --hwe 1E-15. For 1,260 samples proteomics data was available. These samples were merged into a single genotype file and further filtered using plink2 with the options --maf 0.05 --hwe 1E-6 --geno 0.02, leaving data for 5,461,287 genetic variants. The first ten genetic principal components were then computed using plink2 with the --pca option. QMDiab samples were genotyped using the Illumina Omni 2.5 array (version 8) and imputed using the SHAPEIT software with 1000 Genomes (phase3) haplotypes. Genotyping data was available for 325 of the 345 samples with proteomics data.

Proteomics. The workflow used for the proteomics analyses of the Tarkin and the QMDiab samples were essentially identical and have been described in detail before [15]. Briefly, plasma samples were prepared using the Proteograph workflow [19,20] (*Seer, Inc.*) to generate purified peptides that were then analyzed using a dia-PASEF method [26] on a timsTOF Pro 2 mass spectrometer *(Bruker Daltonics)*. Each study was conducted at independent times using two mass spectrometers. One mass spectrometer coincidentally was reused in both studies. DIA-NN (1.8.1) [27] was used to derive peptide and protein intensities. A library-free search based on UniProt version UP000005640_9606 was used, processing the data a second time using the match between runs (MBR) option. Two additional libraries were created, one excluding common (MAF > 10%) protein-altering variant (PAV) containing peptides, and one injecting the alternate alleles into the reference protein sequences. These libraries were referred to as the *PAV-exclusive* library and the *PAV-inclusive* library, respectively. Details of this library generation process have been described elsewhere [15]. DIA-NN's normalized intensities (PG.Normalised) were used as protein readout. 5,753 unique protein groups were quantified, 4,109 were detected in at least 20% of the samples, and 1,980 had less than 20% of missing values. Wherever a protein was detected at this level in more than one nanoparticle run, the one with the largest sum of protein intensities was retained. Protein levels were then log-scaled, residualized using age, sex and the ten first genotype principal components, and finally inverse-normal scaled.

Genome-wide association. Genetic associations of 1,980 residualized and inverse-normal scaled protein levels with 5,461,287 genetic variants in 1,260 samples were evaluated using linear models (plink v1.90b7.1, option --linear). Missing data points were excluded. The Bonferroni significance level for protein associations was $p_{Bonf} = 5x10^{-8}/1,980 = 2.5x10^{-11}$. For six proteins, the residualization did not entirely remove association with these confounders. Three of these proteins presented with inflated GWAS statistics and the corresponding pQTLs were removed (UniProt V9GYE7, B1AKG0, and O15230). Protein associations on correlated variants were clumped on a per-trait basis using an LD cut-off of r^2=0.1 For each cluster the trait sentinel association was identified. Sentinel associations were then clumped using an LD cut-off of r^2=0.9 into loci.

Replication. Replication was attempted using data from 325 samples of the QMDiab study that had joint genotype and proteomics information. Power to replicate 80% of the pQTLs was determined as the 80% quantile of the p-values obtained from 1,000 random samples from Tarkin data set using the number of samples available in the QMDiab for the respective genotype-protein pair.

Overlap with previous OLINK and SOMAscan pQTLs. Summary statistics from UKB-PPP Olink study [14] and deCODE SOMAscan study [13] were downloaded from the respective sites. Associations were reported for 4,660 proteins by deCODE and for 2,908 proteins by UKB-PPP. All protein associations on variants matching one of the Proteograph pQTLs were retrieved. An association reported by deCODE and Olink was considered as significant at levels of $p < 0.05/139/4660$ and $p < 0.05/117/2908$, respectively, accounting for the number of evaluated variants and proteins on the panel.

Evaluation of age and sex associations. Summary statistics for age and sex associations were retrieved from the from the supplementary Excel files of Ferkingstad *et al.* [13] for deCODE SOMAscan (sheet ST01) and of Sun *et al.* [14] for UKB-PPP Olink (sheet ST5). Associations with age and sex for the Tarkin study were computed with PLINK using identical datasets and models as for the genome-wide association (option --linear no-snp). Proteins were matched using UniProt identifiers. In rare cases when there were matches to multiple protein groups, the strongest association was retained.

Evaluation of OLINK and SOMAscan cis-pQTLs. Cis-pQTLs were obtained from the supplementary tables of the respective studies (ST02 from Ferkingstad *et al*. [13] for deCODE SOMAscan and ST10 from Sun *et al*. [14] for UKB-PPP OLINK). pQTLs were limited to variants that were located on autologous chromosomes, had a suitable replication SNP available in both, Tarkin and QMDiab (LD $r^2 > 0.8$), and a minor allele frequency greater than 5% in the Tarkin study. Matching of proteins between platforms was done using Uniprot IDs, allowing for matches to protein groups that contained the Uniprot ID. Ambiguous cases where more than one matching protein group was found were omitted. In cases where protein readouts for multiple nanoparticle runs where available, the one with the highest single number of peptide detections was used. Cases where the number of quantified proteins in Tarkin was below 80% were excluded.

Annotation of pQTLs. snipa.org [28], phenoscanner.medschl.cam.ac.uk [21] and omicsciences.org [29] were used to annotate pQTLs with overlapping information on disease GWAS, gene expression and metabolomics QTLs. Protein epitope changing variants were identified using Ensembl [22].

DATA AVAILABILITY STATEMENT

The MS-proteomics data for the Tarkin study has been deposited with ProteomeXchange under project accession code PXD048709 and will be made public at time of publication. The MS-proteomics data of QMDiab is already available on ProteomeXchange with identifier PXD042852. Summary statistics of the GWAS will be deposited in the GWAS catalog. Consent obtained from the QMDiab study participants does not allow deposition of genetic information in public databases. Researcher affiliated with a research institution may request access to genetic data on an individual basis from the corresponding author (Karsten Suhre, Weill Cornell Medicine - Qatar, Doha, Qatar). Access is subject to approval by the institutional research board of Weill Cornell Medicine - Qatar.

ACKNOWLEDGEMENTS

We are grateful to all study participants for providing their time and blood. K.S. and F.S. are supported by the Biomedical Research Program at Weill Cornell Medicine in Qatar, a program funded by the Qatar

Foundation. K.S. is also supported by Qatar National Research Fund (QNRF) grants NPRP11C-0115-180010 and ARG01-0420-230007. J.L-S. is supported by the National Institute of Health grants R01HL123915, R01HL155742, and U19AI168643. The statements made herein are solely the responsibility of the authors.

We thank Guhan Venkataraman, Amir Alavi and Jian Wang for developing software workflows that were used in the proteogenomic analyses in this work.

AUTHOR CONTRIBUTION STATEMENT

Financial Support: K.S., J.L-S

Study design: K.S., J.L-S.

Data analysis: K.S., H.G., S.B.

Provided Materials and Conducted Experiments: Q.C., A.H., K.M., A.D., N.S., G.T., H.S., V.B.D., F.S.

Manuscript writing: K.S., S.B., J.L-S.

All authors contributed to the interpretation of the results and critically reviewed the manuscript.

COMPETING INTERESTS STATEMENT

H.G. and S.B. are employees and/or stockholders of Seer, Inc.; J.L-S. is a scientific advisor to Precion Inc. and TruDiagnostic; J.L-S. has a sponsored research agreement with TruDiagnostic. J.L-S. previously consulted for Cambrian and Ahara. The other authors declare no competing interests.

REFERENCES

1. Suhre, K., McCarthy, M.I., and Schwenk, J.M. (2021). Genetics meets proteomics: perspectives for large population-based studies. Nature reviews. Genetics *22*, 19-37. 10.1038/s41576-020-0268-2.
2. Plenge, R.M., Scolnick, E.M., and Altshuler, D. (2013). Validating therapeutic targets through human genetics. Nature reviews. Drug discovery *12*, 581-594. 10.1038/nrd4051.
3. 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. Nature communications *8*, 14357. 10.1038/ncomms14357.

4. 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. 10.1038/s41586-018-0175-2.
5. Lourdusamy, A., Newhouse, S., Lunnon, K., Proitsi, P., Powell, J., Hodges, A., Nelson, S.K., Stewart, A., Williams, S., Kloszewska, I., et al. (2012). Identification of cis-regulatory variation influencing protein abundance levels in human plasma. Human molecular genetics *21*, 3719-3726. 10.1093/hmg/dds186.
6. Enroth, S., Johansson, A., Enroth, S.B., and Gyllensten, U. (2014). Strong effects of genetic and lifestyle factors on biomarker variation and use of personalized cutoffs. Nature communications *5*, 4684. 10.1038/ncomms5684.
7. 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 (New York, N.Y.) *361*, 769-773. 10.1126/science.aaq1327.
8. 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 (New York, N.Y.) *374*, eabj1541. doi:10.1126/science.abj1541.
9. Gudjonsson, A., Gudmundsdottir, V., Axelsson, G.T., Gudmundsson, E.F., Jonsson, B.G., Launer, L.J., Lamb, J.R., Jennings, L.L., Aspelund, T., Emilsson, V., and Gudnason, V. (2022). A genome-wide association study of serum proteins reveals shared loci with common diseases. Nature communications *13*, 480. 10.1038/s41467-021-27850-z.
10. Zhao, J.H., Stacey, D., Eriksson, N., Macdonald-Dunlop, E., Hedman, Å.K., Kalnapenkis, A., Enroth, S., Cozzetto, D., Digby-Bell, J., Marten, J., et al. (2023). Genetics of circulating inflammatory proteins identifies drivers of immune-mediated disease risk and therapeutic targets. Nature Immunology *24*, 1540-1551. 10.1038/s41590-023-01588-w.
11. Carland, C., Png, G., Malarstig, A., Kho, P.F., Gustafsson, S., Michaelsson, K., Lind, L., Tsafantakis, E., Karaleftheri, M., Dedoussis, G., et al. (2023). Proteomic analysis of 92 circulating proteins and their effects in cardiometabolic diseases. Clinical proteomics *20*, 31. 10.1186/s12014-023-09421-0.
12. 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. Nature metabolism *2*, 1135-1148. 10.1038/s42255-020-00287-2.
13. 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. Nature genetics *53*, 1712-1721. 10.1038/s41588-021-00978-w.
14. Sun, B.B., Chiou, J., Traylor, M., Benner, C., Hsu, Y.-H., Richardson, T.G., Surendran, P., Mahajan, A., Robins, C., Vasquez-Grinnell, S.G., et al. (2023). Plasma proteomic associations with genetics and health in the UK Biobank. Nature. 10.1038/s41586-023-06592-6.
15. Suhre, K., Venkataraman, G.R., Guturu, H., Halama, A., Stephan, N., Thareja, G., Sarwath, H., Motamedchaboki, K., Donovan, M.K.R., Siddiqui, A., et al. (2024). Nanoparticle enrichment mass-spectrometry proteomics identifies protein-altering variants for precise pQTL mapping. Nature communications *15*, 989. 10.1038/s41467-024-45233-y.
16. Johansson, Å., Enroth, S., Palmblad, M., Deelder, A.M., Bergquist, J., and Gyllensten, U. (2013). Identification of genetic variants influencing the human plasma proteome. Proceedings of the National Academy of Sciences of the United States of America *110*, 4673-4678. 10.1073/pnas.1217238110.

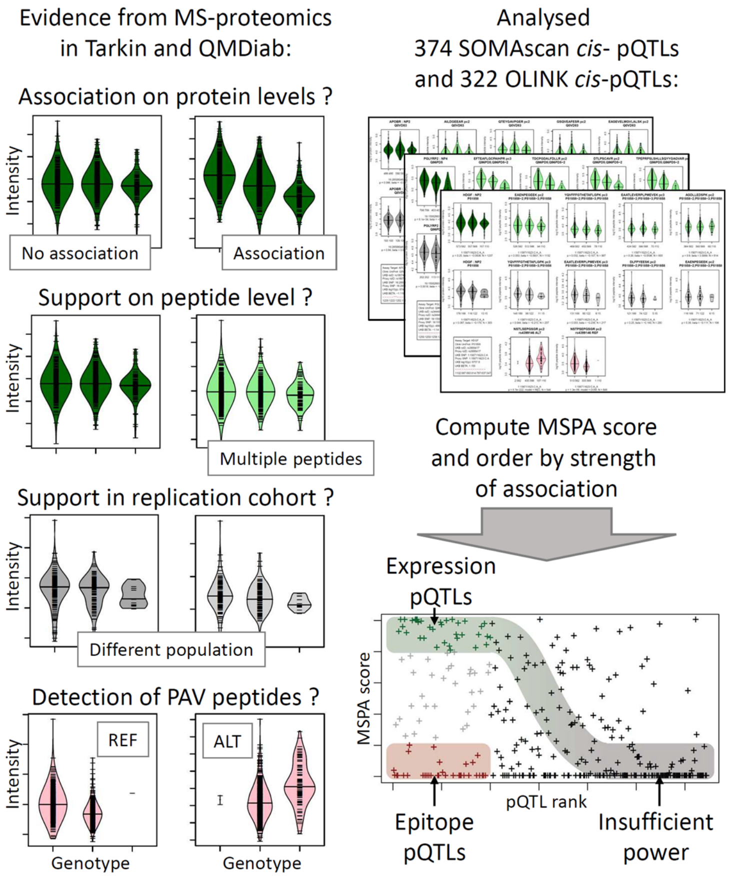
17. Xu, F., Yu, E.Y.-W., Cai, X., Yue, L., Jing, L.-p., Liang, X., Fu, Y., Miao, Z., Yang, M., Shuai, M., et al. (2023). Genome-wide genotype-serum proteome mapping provides insights into the cross-ancestry differences in cardiometabolic disease susceptibility. Nature communications *14*, 896. 10.1038/s41467-023-36491-3.
18. Niu, L., Stinson, S.E., Holm, L.A., Lund, M.A.V., Fonvig, C.E., Cobuccio, L., Meisner, J., Juel, H.B., Thiele, M., Krag, A., et al. (2023). Plasma Proteome Variation and its Genetic Determinants in Children and Adolescents. medRxiv : the preprint server for health sciences, 2023.2003.2031.23287853. 10.1101/2023.03.31.23287853.
19. Liu, Y., Wang, J., Xiong, Q., Hornburg, D., Tao, W., and Farokhzad, O.C. (2021). Nano-Bio Interactions in Cancer: From Therapeutics Delivery to Early Detection. Accounts of chemical research *54*, 291-301. 10.1021/acs.accounts.0c00413.
20. Ferdosi, S., Stukalov, A., Hasan, M., Tangeysh, B., Brown, T.R., Wang, T., Elgierari, E.M., Zhao, X., Huang, Y., Alavi, A., et al. (2022). Enhanced Competition at the Nano-Bio Interface Enables Comprehensive Characterization of Protein Corona Dynamics and Deep Coverage of Proteomes. Advanced materials (Deerfield Beach, Fla.) *34*, e2206008. 10.1002/adma.202206008.
21. Kamat, M.A., Blackshaw, J.A., Young, R., Surendran, P., Burgess, S., Danesh, J., Butterworth, A.S., and Staley, J.R. (2019). PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. Bioinformatics (Oxford, England) *35*, 4851-4853. 10.1093/bioinformatics/btz469.
22. Martin, F.J., Amode, M.R., Aneja, A., Austine-Orimoloye, O., Azov, Andrey G., Barnes, I., Becker, A., Bennett, R., Berry, A., Bhai, J., et al. (2022). Ensembl 2023. Nucleic Acids Research *51*, D933-D941. 10.1093/nar/gkac958.
23. Chen, Q., Dwaraka, V.B., Carreras-Gallo, N., Mendez, K., Chen, Y., Begum, S., Kachroo, P., Prince, N., Went, H., Mendez, T., et al. (2023). OMICmAge: An integrative multi-omics approach to quantify biological age with electronic medical records. bioRxiv, 2023.2010.2016.562114. 10.1101/2023.10.16.562114.
24. Mook-Kanamori, D.O., Selim, M.M., Takiddin, A.H., Al-Homsi, H., Al-Mahmoud, K.A., Al-Obaidli, A., Zirie, M.A., Rowe, J., Yousri, N.A., Karoly, E.D., et al. (2014). 1,5-Anhydroglucitol in saliva is a noninvasive marker of short-term glycemic control. The Journal of clinical endocrinology and metabolism *99*, E479-483. 10.1210/jc.2013-3596.
25. Yousri, N.A., Mook-Kanamori, D.O., Selim, M.M., Takiddin, A.H., Al-Homsi, H., Al-Mahmoud, K.A., Karoly, E.D., Krumsiek, J., Do, K.T., Neumaier, U., et al. (2015). A systems view of type 2 diabetes-associated metabolic perturbations in saliva, blood and urine at different timescales of glycaemic control. Diabetologia *58*, 1855-1867. 10.1007/s00125-015-3636-2.
26. Meier, F., Brunner, A.-D., Frank, M., Ha, A., Bludau, I., Voytik, E., Kaspar-Schoenefeld, S., Lubeck, M., Raether, O., Bache, N., et al. (2020). diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition. Nature Methods *17*, 1229-1236. 10.1038/s41592-020-00998-0.
27. Demichev, V., Szyrwiel, L., Yu, F., Teo, G.C., Rosenberger, G., Niewienda, A., Ludwig, D., Decker, J., Kaspar-Schoenefeld, S., Lilley, K.S., et al. (2022). dia-PASEF data analysis using FragPipe and DIA-NN for deep proteomics of low sample amounts. Nature communications *13*, 3944. 10.1038/s41467-022-31492-0.
28. Arnold, M., Raffler, J., Pfeufer, A., Suhre, K., and Kastenmüller, G. (2015). SNiPA: an interactive, genetic variant-centered annotation browser. Bioinformatics (Oxford, England) *31*, 1334-1336. 10.1093/bioinformatics/btu779.
29. 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 (New York, N.Y.) *374*, eabj1541. 10.1126/science.abj1541.

TABLES

Table 1: List of 30 previously unreported replicated pQTLs discovered using the Seer Proteograph platform.

Gene	UniProt	rsID	SNP	Eff	EAF	Discovery (Tarkin)			Replication (QMDiab)			Gene locus
						N	Beta	Log10p	N	Beta	Log10p	
LEFTY1	O75610	rs360057	1:226074563:T:G	G	35.2%	1255	-0.507	36.0	322	-0.300	4.2	cis
ANGPTL6	Q8NI99	rs6542680	2:3640142:C:T	C	17.3%	1257	0.343	11.2	325	0.288	3.9	COLEC11
C2orf40	Q9H1Z8	rs13014521	2:106687456:G:C	C	11.5%	1214	-0.809	37.2	316	-0.795	8.3	cis
MAN1C1	Q9NR34	rs4305381	3:126249877:A:C	C	24.5%	1123	-0.420	17.1	319	-0.542	6.5	C3orf22,CHST13
AMBP	P02760	rs1056522	3:126261345:G:A	A	30.6%	1257	-0.713	67.4	325	-0.514	7.8	CHST13
CCDC132	Q96JG6	rs1056522	3:126261345:G:A	A	30.6%	1250	-0.665	57.2	257	-0.575	7.1	CHST13
ZNF618	Q5T7W0	rs9835865	3:186380167:A:T	T	49.2%	1153	0.421	23.6	279	0.339	4.6	FETUB,HRG
ZNF618	Q5T7W0	rs11708856	3:186390393:G:C	C	6.4%	1153	-0.682	12.9	279	-0.795	6.9	FETUB,HRG
GNB2	P62879	rs1042464	3:186395572:A:T	A	49.1%	1252	-0.310	14.2	320	-0.336	4.8	HRG
CTSS	P25774	rs5030062	3:186454180:A:C	C	38.9%	1255	0.313	14.4	325	0.372	5.9	KNG1
HLA-G	Q5RJ85	rs2517718	6:29916391:A:C	C	43.1%	1213	-0.658	65.5	315	-0.578	12.0	cis
HLA-G	Q5RJ85	rs2442717	6:31322470:C:G	C	16.7%	1213	0.747	50.2	315	0.590	3.9	cis
FUCA2	Q9BTY2	rs11155297	6:143825104:G:T	T	25.1%	1253	-0.523	30.1	325	-0.651	11.2	cis
SRCRB4D	Q8WTU2	rs111666614	7:76044757:G:T	T	7.1%	1140	0.528	11.0	318	0.725	7.5	cis
PEBP4	Q96S96	rs3087803	8:22570901:C:T	T	9.0%	1255	0.925	46.6	325	0.987	11.4	cis
PEBP4	Q96S96	rs1129474	8:22584718:T:C	C	49.3%	1255	-0.454	30.7	325	-0.529	11.6	cis
CD34	P28906	rs687621	9:136137065:A:G	G	34.4%	1051	0.357	14.5	269	0.374	5.4	ABO
MMRN2	Q9H8L6	rs34587013	10:88696622:C:G	G	8.3%	1250	-0.493	11.6	325	-0.510	4.0	cis
TSKU	Q8WUA8	rs1149596	11:76469093:C:T	T	13.9%	1253	-0.477	16.5	321	-0.353	3.9	cis
FXYD2	P54710	rs4936409	11:117694392:A:G	G	48.3%	1237	-0.406	24.2	322	-0.365	6.2	cis
PRB1	A0A4W8X8U3	rs7966710	12:11522616:G:A	G	26.7%	1133	-0.609	37.4	238	-0.462	4.2	cis
PROZ	P22891	rs559054	13:113800622:T:C	T	37.6%	1256	-0.383	21.2	325	-0.385	6.7	cis
GALC	P54803	rs380142	14:88393918:A:C	C	45.2%	1252	-1.022	205.0	317	-0.889	40.6	cis
GALC	P54803	rs12434101	14:88478827:T:C	C	8.8%	1252	-0.625	18.9	317	-0.616	8.8	cis
GALC	P54803	rs12323470	14:88496080:G:T	T	13.9%	1252	0.700	36.3	317	0.866	12.4	cis
IGHV2-70	A0A0C4DH43	rs10134517	14:107173745:T:C	C	29.6%	1090	0.504	27.9	283	0.516	8.0	cis
IGHV2-70	A0A0C4DH43	rs7161739	14:107184357:G:A	A	13.7%	1090	-0.562	15.8	283	-0.620	8.4	cis
FN3K	Q9H479	rs3848403	17:80693899:C:T	T	50.0%	1243	-0.673	69.2	281	-0.412	5.9	cis
RCN3	Q96D15	rs73582463	19:50037446:C:G	G	8.4%	1255	0.858	35.5	324	0.549	5.0	cis
BPIFA1	Q9NP55	rs6059187	20:31828265:A:G	G	49.1%	1245	-0.388	23.9	323	-0.377	6.9	cis

GRAPHICAL ABSTRACT summarizing the approach taken to identify potential epitope effects.



FIGURES

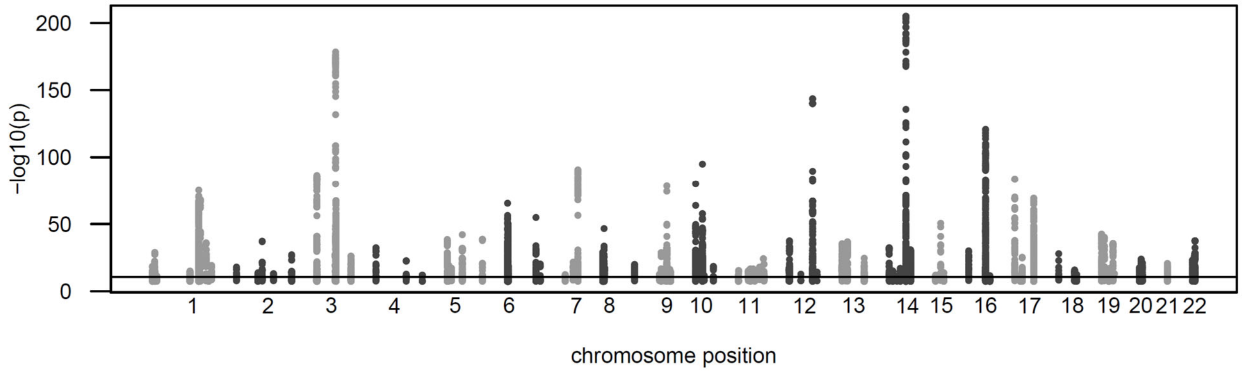
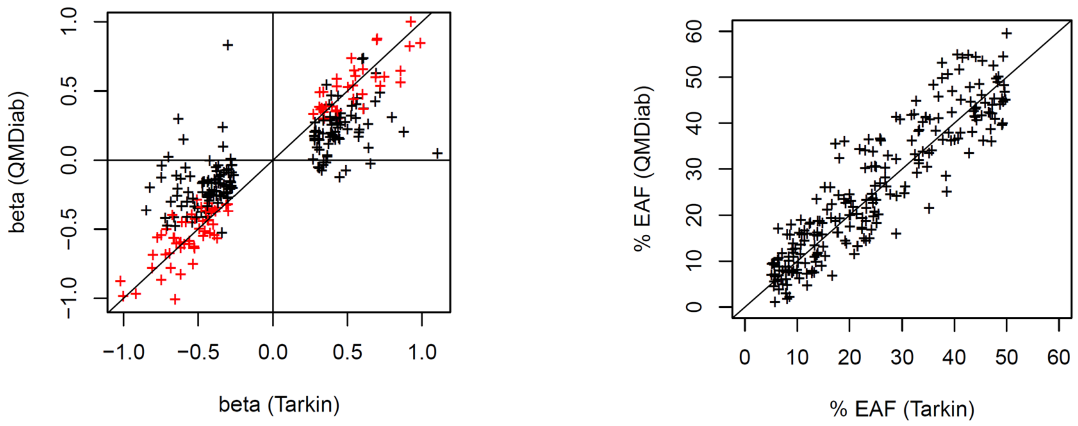


Figure 1: Manhattan plot. Shown are all protein associations that reached a significance level $p < 5x10^{-8}$ in the discovery study.



(A)

(B)

Figure 2: Effect size and effect allele frequencies. Scatterplot of the pQTL effect sizes from the discovery (Tarkin) and the replication (QMDiab) study, replicated loci are shown in red (A); Scatterplot of the effect allele frequencies (EAF), Tarkin versus QMDiab (B).

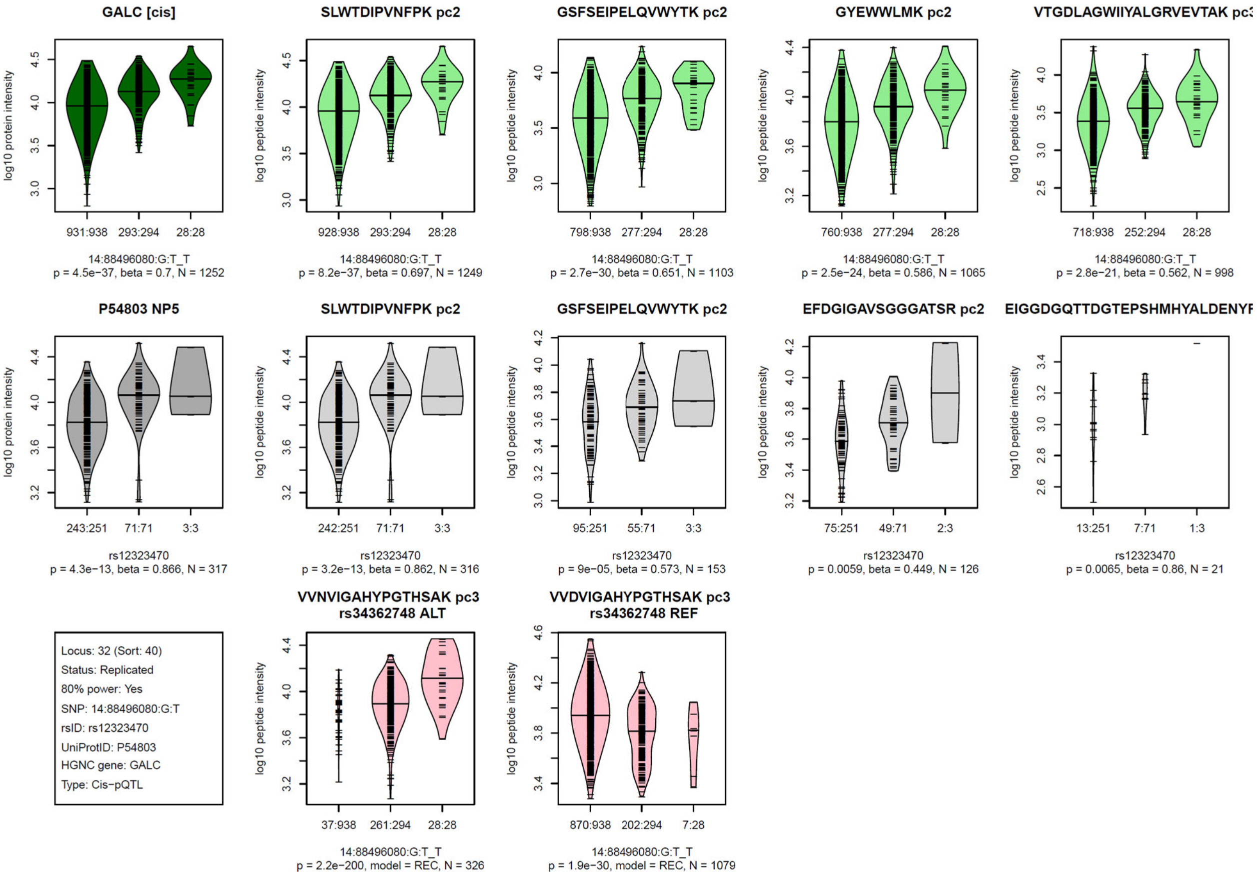


Figure 3: Visualization of the Proteograph MS-proteomics data for a prototypical pQTL. Violin plots of log-scaled engine-normalized protein and peptide intensities by genotype for the indicated protein and genetic variant (bottom left box, details are in **Table S2**); Top row (green): Tarkin data, using *PAV-exclusive* library, Middle row (grey): QMDiab data, using *PAV-exclusive* library, Bottom row (red): Tarkin data, using *PAV-inclusive* library, limited to PAV containing peptides; Plot titles indicate the protein UniProt identifier and the pQTL type (*cis*/*trans*) above the Tarkin protein plot, and HGNC gene name and nanoparticle number are above the QMDiab protein plot; Titles of the peptide plots indicate the respective peptide sequences and precursor charges (pc), and additionally the PAV variant rsID and allele type (REF/ALT) for the PAV peptide plots; Summary statistics are reported below the plots based on linear models with residualized and inverse-normal scaled data for associations with the *PAV-exclusive* library data (top two rows) and Fisher's exact test for *PAV-inclusive* library data (bottom row); Associations with peptide data are sorted by increasing p-values from left to right and limited to a maximum of four plots; Whenever peptides were detected at multiple precursor charge values, only the strongest association was plotted; Numbers at the x-axis tick marks indicate the number of detected peptides by genotype followed by the number a samples with the corresponding genotype (e.g. 662:665); Genotypes are ordered as (1) other allele, (2) heterozygote, (3) effect allele, where the effect allele is indicated following the SNP name (chr:pos:ref:alt_eff, e.g. 3:49721532:G:A_A). Similar plots for all 252 pQTLs are provided as **Figure S2**.

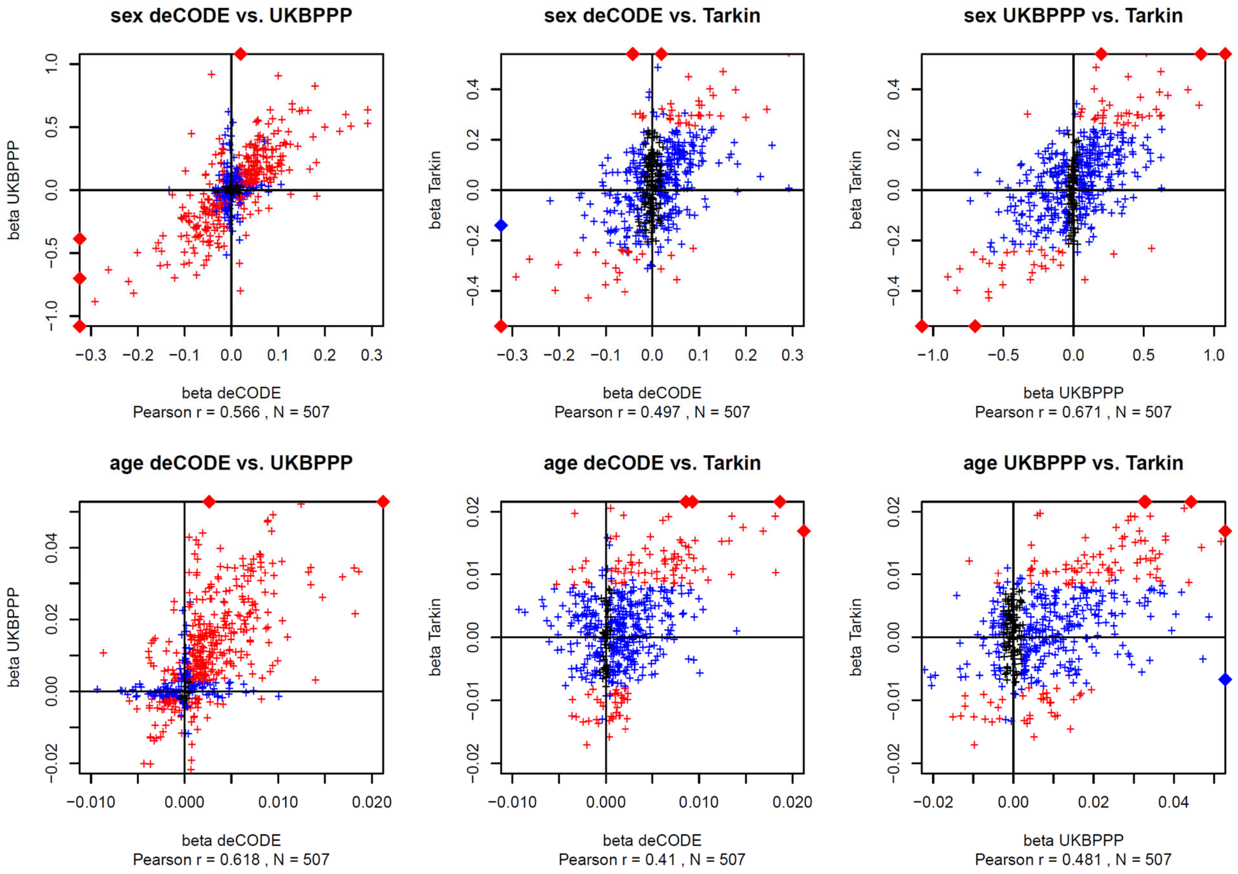


Figure 4: Scatterplot of the effect sizes for the protein associations with sex and age. Summary statistics for the associations with affinity proteomics were from the respective GWAS studies. Associations that reached Bonferroni significance in both respective studies are in red and in one study are in blue ($p < 0.05$ / number of reported associations). The effect sizes (beta) are reported in units of standard deviations (s.d.). Data points outside the plotting window are indicated by diamonds on the plot frames. Plot data are available in **Table ST4**. Scatterplots are limited to 507 unique proteins that were reported by all three studies. In cases where data for multiple affinity binders was reported, the most significant association was retained.

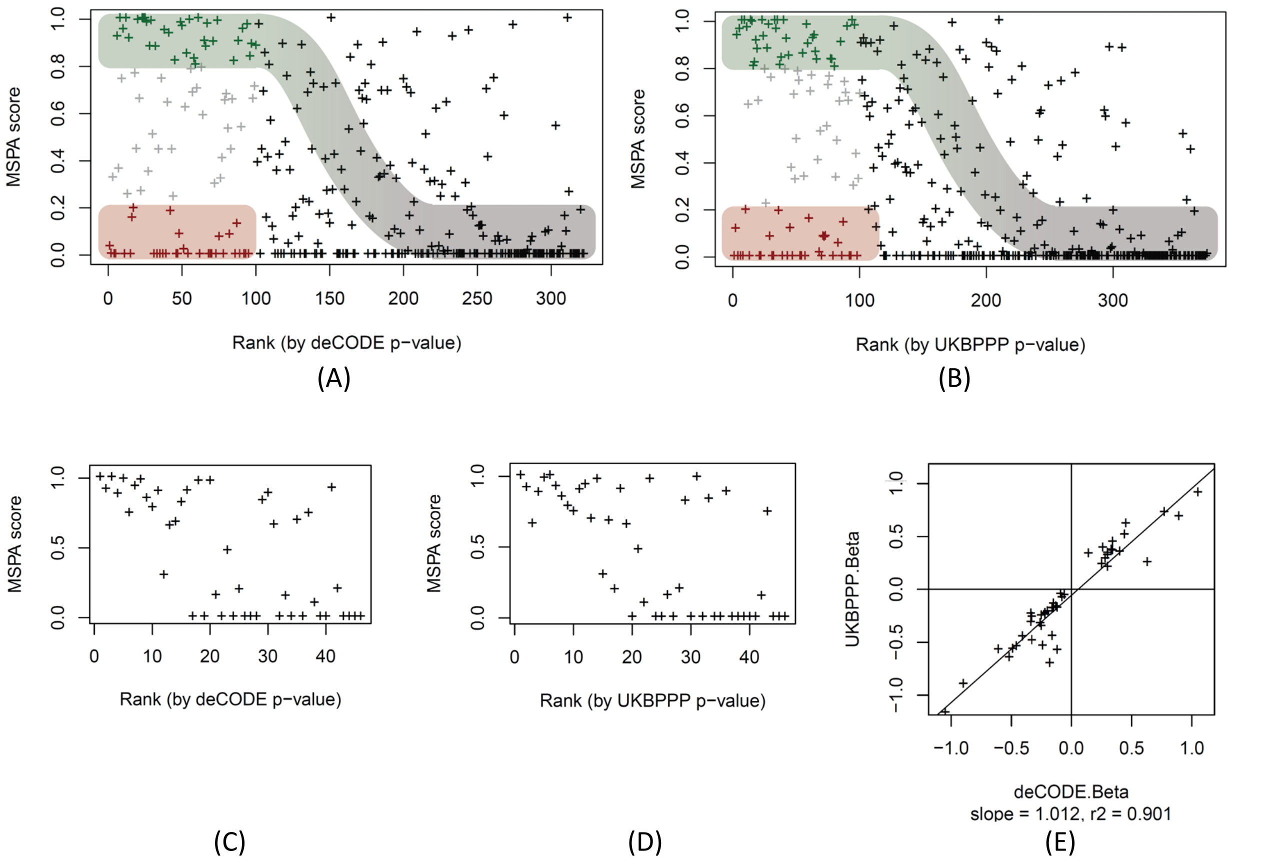


Figure 5: MS-peptide association score plotted by pQTL rank. Scatterplot of the MSPA scores against the rank of the affinity proteomics pQTLs of the deCODE SOMAscan (panel A, data in **Table S7**) and the UKB-PPP OLINK (panel B, data in **Table S8**) studies, ranked starting with the lowest p-value. The first 100 pQTLs (out of 322 pQTLs for SOMAscan and 374 pQTLs for Olink) are coloured to indicate likely protein expression QTLs (MSPA score > 0.8; green) and likely epitope effect driven pQTLs (MSPA score < 0.2; red), the sigmoid curve indicates the assumed dependence of power to detect a pQTL as a function of the strength of the association, approximated by the rank of the pQTL in the respective study; MSPA scores limited to 46 pQTLs that were reported on the same variant in deCODE (panel C) and UKBPPP (panel D); scatterplot of the effect size (beta) of the 46 pQTLs reported deCODE and UKBPPP (panel E, data in **Table S9**).

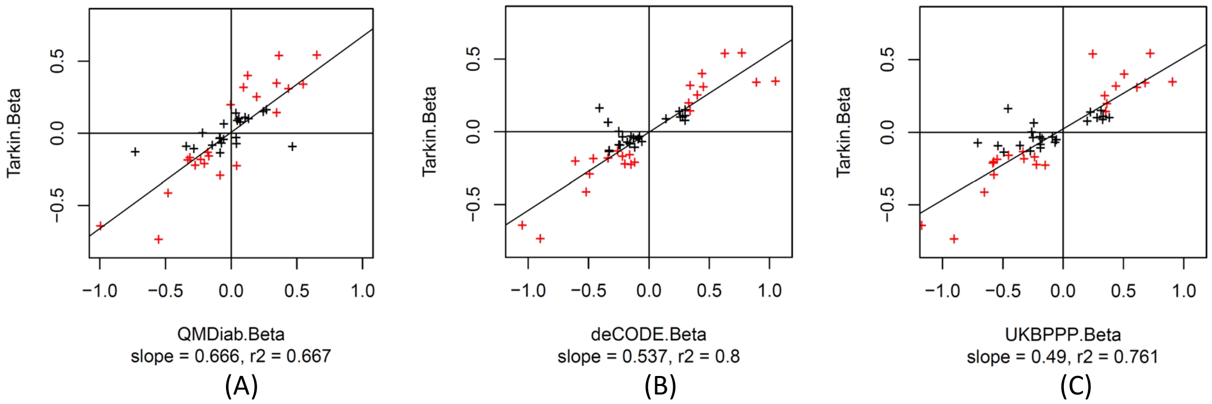


Figure 6: Scatterplot of the effect sizes (beta) of 46 pQTLs that were reported by deCODE and UKBPPP on the same SNP and for which association data was also available for Tarkin and QMDiab; Tarkin vs. QMDiab (A), Tarkin vs. deCODE (B), Tarkin vs. UKBPPP (C); associations that reach a significance level of $p < 0.05 / 46$ in Tarkin are in red (**Table S9**).

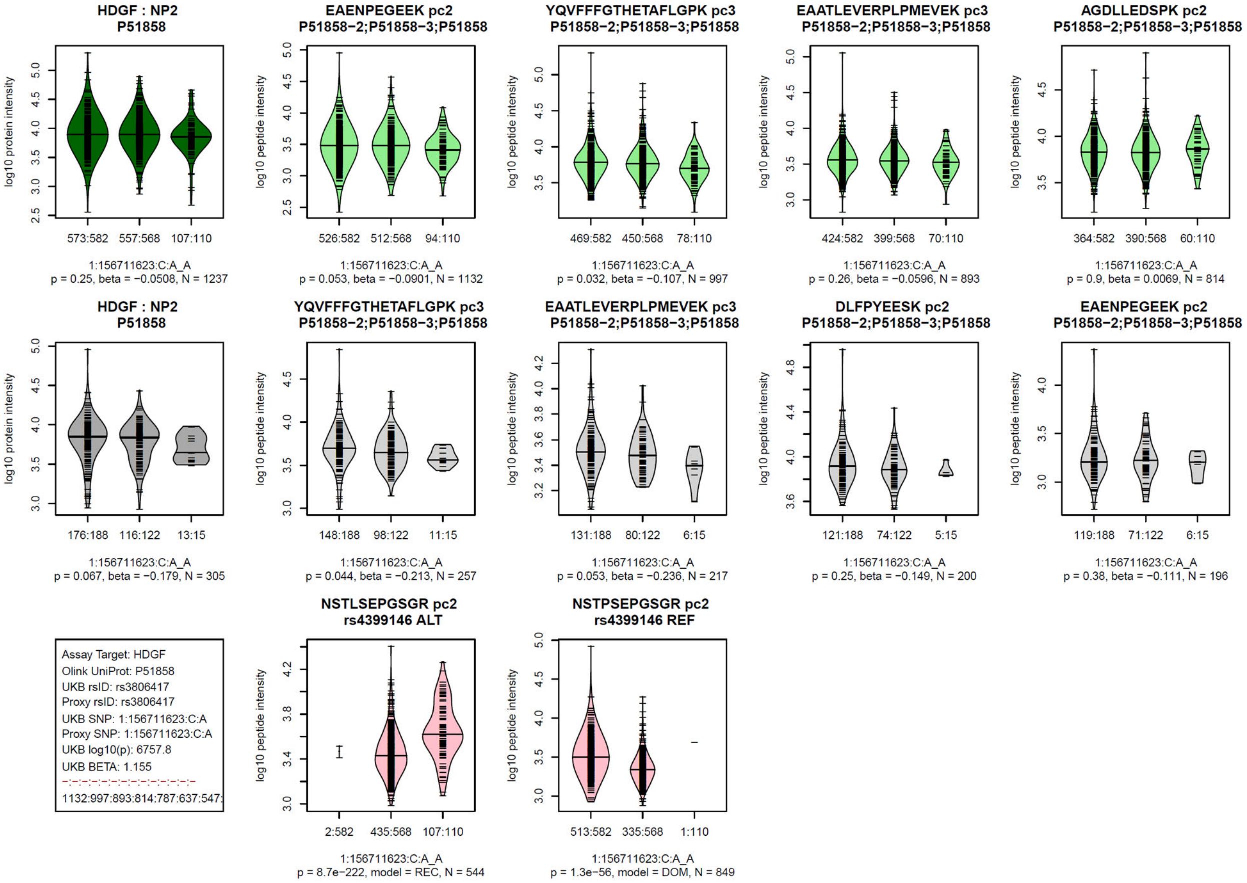
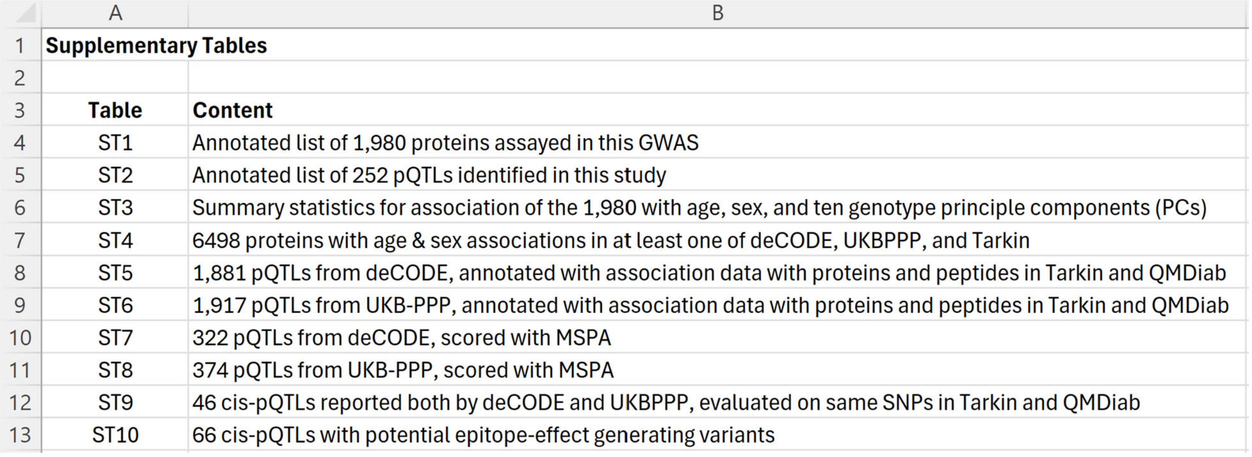


Figure 7: Example of a pQTL that is likely affected by an epitope effect. See legend of **Figure 3** for legend; similar plots for all 374 pQTLs with OLINK data and for all 322 pQTLs with SOMAscan data are provided as **Figures S3** and **S4**; data is in **Tables S6** and **S7**.

SUPPLEMENTARY MATERIAL

Supplementary Tables [provided as a multi-sheet MS Excel file]



	A	B
1	**Supplementary Tables**	
2		
3	**Table**	**Content**
4	ST1	Annotated list of 1,980 proteins assayed in this GWAS
5	ST2	Annotated list of 252 pQTLs identified in this study
6	ST3	Summary statistics for association of the 1,980 with age, sex, and ten genotype principle components (PCs)
7	ST4	6498 proteins with age & sex associations in at least one of deCODE, UKBPPP, and Tarkin
8	ST5	1,881 pQTLs from deCODE, annotated with association data with proteins and peptides in Tarkin and QMDiab
9	ST6	1,917 pQTLs from UKB-PPP, annotated with association data with proteins and peptides in Tarkin and QMDiab
10	ST7	322 pQTLs from deCODE, scored with MSPA
11	ST8	374 pQTLs from UKB-PPP, scored with MSPA
12	ST9	46 cis-pQTLs reported both by deCODE and UKBPPP, evaluated on same SNPs in Tarkin and QMDiab
13	ST10	66 cis-pQTLs with potential epitope-effect generating variants

Supplementary Figures

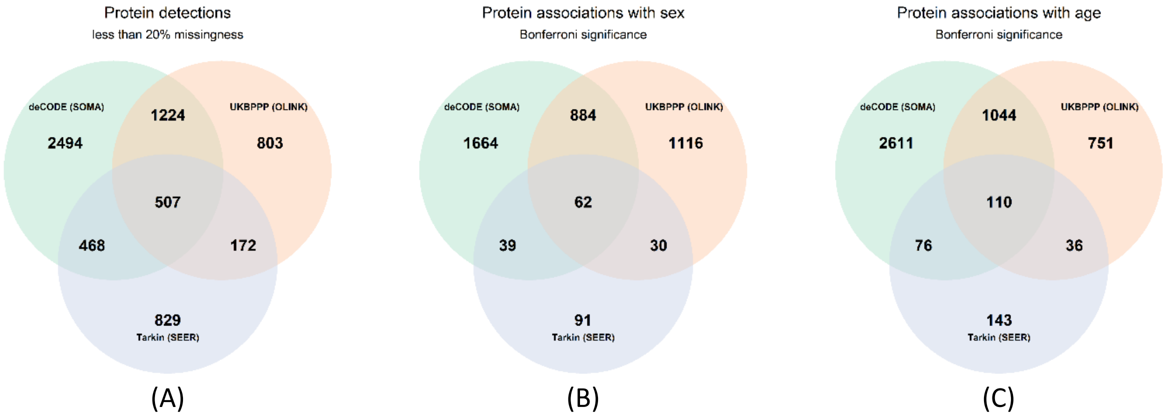


Figure S1: Venn diagrams. Proteins reported by the three studies, 1,980 proteins with <20% missingness for Tarkin, out of 5,753 proteins quantified across the discovery cohort (A); Proteins associated with sex (B) and age (C) at a significance level of 0.05 / number of proteins reported.

The following Figures are provided as separate PDF files:

Figure S2: Violin plots of protein and peptide levels by genotype for 252 pQTLs identified in this study. See Figure 3 for legend. File: Figure_S2_PGWAS_replicate_check.20240523.pdf

Figure S3: Violin plots of protein and peptide levels by genotype for 322 pQTLs identified by the deCODE OLINK study. See Figure 7 for legend. File: Figure_S3_PGWAS_test_epitope_SOMA.20240523.pdf

Figure S4: Violin plots of protein and peptide levels by genotype for 374 pQTLs identified by the UKB-PPP SOMAscan study. See Figure 7 for legend. File: Figure_S4_PGWAS_test_epitope_OLINK.20240523.pdf