Nanoparticle Enrichment Mass-Spectrometry Proteomics Identifies Protein Altering Variants for Precise pQTL Mapping

Karsten Suhre[1,2,*], *Guhan Ram Venkataraman*[3], *Harendra Guturu*[3], *Anna Halama*[1,2], *Nisha Stephan*[1], *Gaurav Thareja*[1], *Hina Sarwath*[4], *Khatereh Motamedchaboki*[3], *Margaret Donovan*[3], *Asim Siddiqui*[3], *Serafim Batzoglou*[3], *Frank Schmidt*[4]

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

[2] Department of Biophysics and Physiology, Weill Cornell Medicine, New York, NY 10065, U.S.A.

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

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

[*] Correspondence to K.S. (kas2049@qatar-med.cornell.edu)

ABSTRACT

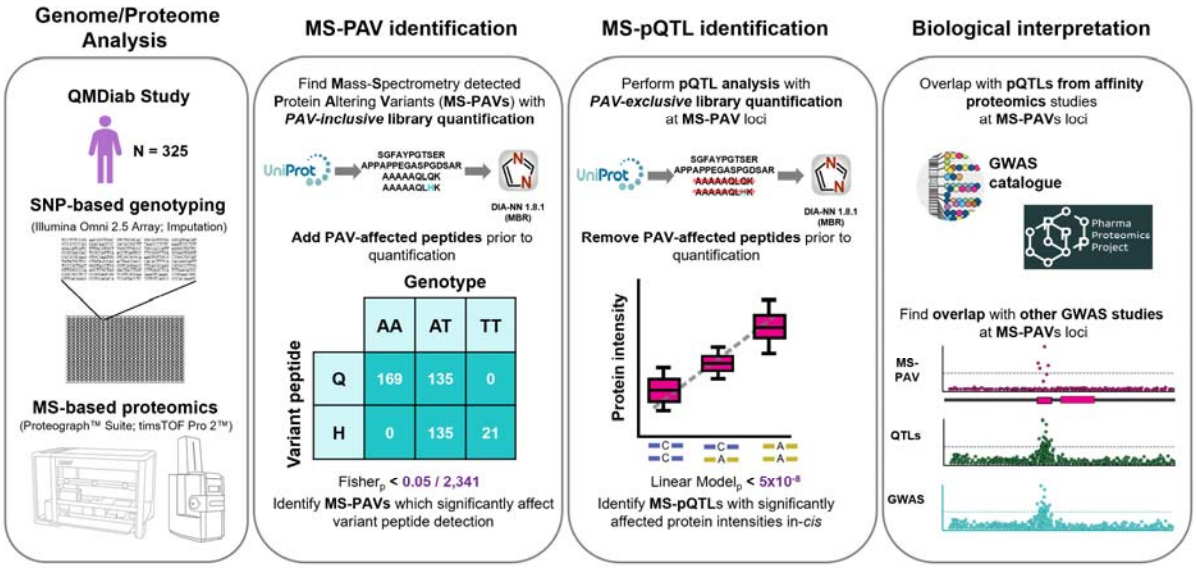
Genome-wide association studies (GWAS) with proteomics generate hypotheses on protein function and offer genetic evidence for drug target prioritization. Although most protein quantitative loci (pQTLs) have so far been identified by high-throughput affinity proteomics platforms, these methods also have some limitations, such as uncertainty about target identity, non-specific binding of aptamers, and inability to handle epitope-modifying variants that affect affinity binding. Mass spectrometry (MS) proteomics has the potential to overcome these challenges and broaden the scope of pQTL studies. Here, we employ the recently developed MS-based Proteograph™ workflow (*Seer, Inc.*) to quantify over 18,000 unique peptides from almost 3,000 proteins in more than 320 blood samples from a multi-ethnic cohort. We implement a bottom-up MS-proteomics approach for the detection and quantification of blood-circulating proteins in the presence of protein altering variants (PAVs). We identify 184 PAVs located in 137 genes that are significantly associated with their corresponding variant peptides in MS data (MS-PAVs). Half of these MS-PAVs (94) overlap with *cis*-pQTLs previously identified by affinity proteomics pQTL studies, thus confirming the target specificity of the affinity binders. An additional 54 MS-PAVs overlap with *trans*-pQTLs (and not *cis*-pQTLs) in affinity proteomics studies, thus identifying the putatively causal *cis*-encoded protein and providing experimental evidence for its presence in blood. The remaining 36 MS-PAVs have not been previously reported and include proteins that may be inaccessible to affinity proteomics, such as a variant in the incretin pro-peptide (GIP) that associates with type 2 diabetes and cardiovascular disease. Overall, our study introduces a novel approach for analyzing MS-based proteomics data within the GWAS context, provides new insights relevant to genetics-based drug discovery, and highlights the potential of MS-proteomics technologies when applied at population scale.

Keywords: Proteomics, Mass spectrometry, Proteograph™ workflow, Genome-wide association studies, Protein quantitative trait loci, Protein altering variants.

Highlights

- This is the first pQTL study that uses the Proteograph™ (*Seer Inc.*) mass spectrometry-based proteomics workflow.
- We introduce a novel bottom-up proteomics approach that accounts for protein altering variants in the detection of pQTLs.
- We confirm the target and potential epitope effects of affinity binders for *cis*-pQTLs from affinity proteomics studies.
- We establish putatively causal proteins for known affinity proteomics *trans*-pQTLs and confirm their presence in blood.
- We identify novel protein altering variants in proteins of clinical relevance that may not be accessible to affinity proteomics.

Graphical abstract



INTRODUCTION

Large-scale studies of the plasma proteome using extensive biobanks have attracted increasing interest because of their potential to inform drug development by supplementing insights gained from genome-wide disease association studies. Identifying genetic variants associated with protein expression levels (protein quantitative trait loci, or pQTLs) can unveil proteins involved in key biological processes that affect complex traits and disease etiology[1].

The two main technologies employed to quantify protein levels in biological samples are affinity-binding proteomics and mass spectrometry (MS) proteomics. Most large-scale proteomics studies to date have relied on affinity proteomics technologies, utilizing variants of Olink's antibody-based Proximity Extension Assay or Somalogic's aptamer-based SOMAscan platform[2, 3, 4, 5, 6, 7]. The UK Biobank Pharma Proteomics Project (UKB-PPP) recently published initial results from a study that quantified 1,500 proteins in blood plasma from 53,000 UKB participants using the Olink platform[8, 9], and the Fenland study analyzed over 4,700 proteins in more than 10,000 individuals using the SOMAscan technology[10], identifying tens of thousands of pQTLs.

Affinity proteomics approaches can deliver quantitative readouts for hundreds and even thousands of blood-circulating proteins in a high-throughput manner, but also possess certain limitations[1, 11]. Notably, they are exposed to interference from genetic variants that can change the protein's epitope (structure) and modify the antibody or aptamer binding affinities, resulting in ambiguous or erroneous associations[11, 12]. Additionally, establishing target specificity for affinity binders is challenging and must be determined on an individual basis under various physiological conditions. Although the literature considers a genetic association at the gene locus that encodes the protein targeted by a given affinity-binder (*cis*-pQTL) as confirmatory evidence for target specificity, cross-reactivity with other proteins cannot be ruled out in such cases. Some protein classes may also be unsuitable for quantification by affinity binding (e.g., unfolded pro-peptides).

Epitope-modifying variants can result in false-positive associations between genetic variants and protein levels. Additionally, such variants often have a biological impact on the protein function rather than on protein level. A recent study showed that approximately 50% of putative epitope-modifying variants colocalize with GWAS associations, suggesting that these variants modify

protein properties rather than protein abundance[7]. Consequently, genetic epitope effects caused by non-synonymous variation pose a significant challenge to the analysis and application of large-scale affinity proteomics-based pQTL studies for drug development: the effect of an epitope-modifying variant on the outcome might not be through the protein level, and thus, therapeutic changes to the protein level might not yield the desired effect.

MS-based proteomics has the potential to alleviate some of the issues faced by affinity proteomics by directly measuring variant peptides originating from protein altering genetic variants. In a bottom-up MS-proteomics approach, peptides (either generated by *in silico* digestion of a comprehensive protein database or curated experimentally) are matched against mass spectra collected by MS analysis of enzymatically digested protein extracts. Modern mass spectrometers, equipped with liquid chromatography and ion mobility separation capabilities, enable the collection of hundreds of thousands of peptide fragmentation spectra at high mass-resolution in a data independent acquisition (DIA) mode[13]. Such methods can potentially identify genetic epitope effects, as they provide peptide-level sequence readouts. Additionally, they may identify proteins that are not amenable to affinity binding and resolve potentially disease-relevant protein post-translational modifications.

However, bottom-up MS-proteomics approaches also present technological challenges related to peptide and protein identification and quantification[14, 15, 16]. In the context of pQTL studies, one such challenge is quantifying protein levels in the presence of genetic variation[17]. Most current analyses do not account for genetic variation, because incorporating all possible variants would result in a significant increase in spectral library size and false-positive identifications. Consequently, standard proteomic libraries fail to detect variant peptides in homozygous alternate allele carriers and falsely suggest reduced protein levels in heterozygotes, leading to genotype-dependent protein level measurements. This problem is exacerbated by genotype-specific instrumental and technical effects, such as genotype-specific shifts in fragmentation, ionization, ion mobility, and liquid chromatography properties.

Here, we examine these technology-specific challenges and propose potential solutions for effectively utilizing bottom-up MS-based proteomics in conjunction with available genetic variation information. We employ the MS-based Proteograph™ (Seer Inc.) workflow[18, 19] to quantify protein and peptide intensities in blood samples from individuals of a multi-ethnic

cohort. The Proteograph™ workflow uses five physicochemically distinct nanoparticles that each enrich different proteins, thereby compressing the dynamic range of proteins analyzed downstream by DIA-MS[20]. Depending on protein abundance and biophysical properties, some peptides can be detected with two or more of the nanoparticles included in the Proteograph™ Assay. Detections by distinct nanoparticles can be considered technical replicates under different protein extraction protocols and offer additional internal validation of the data.

To account for genetic variability in peptide sequences, we implement a data analysis protocol (see **Methods**) that includes all single nucleotide protein altering variants (PAVs) that are present in the study population at a minor allele frequency (MAF) higher than 10%. Given the size of our cohort, we expect at least 2-3 individuals to be homozygous for the minor allele at this level. We introduce these PAVs into the protein database (UniProt), translate the variants to amino-acid space, and perform *in silico* digestion. We then create three spectral libraries: one where we keep only the peptides that correspond to the reference alleles (termed the "*reference* library"), one where we include both reference and variant peptides (termed the "*PAV-inclusive* library"), and one where we exclude all variant peptides and their respective reference peptides (termed the "*PAV-exclusive* library"). Note that the *reference* library corresponds to what is currently used in standard DIA-MS analyses. Using the three different libraries, we then quantify peptide and protein intensities using DIA-NN[16].

Next, we test the presence of the reference or alternate allele of PAVs for association with the presence or absence of the resulting variant peptide(s) in the proteome of the respective sample donor (detected with the *PAV-inclusive* library) using the Fisher's Exact test. Note that a PAV can give rise to multiple matching variant peptides, including peptides that differ by a single amino acid as well as more complex situations, e.g. when the PAV involves a trypsin cleavage site or a protein modification site. We use the term MS-PAV to refer to a PAV that associates significantly (after correcting for multiple tests) with its matching MS-detected variant peptide(s). We then ask whether the identified MS-PAVs also change the corresponding blood protein intensities. For this purpose, we test for association between the protein intensities (obtained using the *PAV-exclusive* library) with the copy number of the alternate allele of the respective MS-PAV as the dependent variable, as generally practiced in pQTL studies. We use the term MS-pQTL to refer to a PAV that associates with both the detection of the respective variant peptide (using the *PAV-inclusive* library) and the protein intensity (using the *PAV-*

exclusive library). We show that pQTLs that have been identified by large affinity proteomics studies can be characterized by overlapping them with MS-PAVs and MS-pQTLs from MS-proteomics studies.

Our study comprises the following steps: First, we identify MS-PAVs and MS-pQTLs using samples from a multiethnic clinical cohort. Then, we query the summary statistics of the two largest pQTL studies, which used the Olink and the SOMAscan platforms, respectively[9, 10], to evaluate the power of this approach in identifying relevant pQTLs. Finally, we discuss new biological insights derived from this study by overlapping MS-PAVs and MS-pQTLs with GWAS associations with other phenotypes (**Figure 1**).

RESULTS

We identify 184 MS-PAVs by adding protein altering variants to a bottom-up proteomics approach.

Citrate plasma samples were obtained from 345 individuals who participated in the Qatar Metabolomics study of Diabetes (QMDiab)[21, 22]. The previously unthawed samples (aliquot of 240 µL per sample) were processed using the Proteograph™ Product Suite (*Seer, Inc.*)[18, 19] (see **Methods**). Briefly, samples were incubated with five proprietary physicochemically distinct nanoparticles for protein corona formation. Nanoparticle-bound proteins were captured, digested using trypsin, and then analyzed using a dia-PASEF method[13] on a timsTOF Pro 2 mass spectrometer *(Bruker Daltonics)*. All MS files were processed using the DIA-NN software (version 1.8.1) using library-free search with match-between-runs (MBR) enabled against the UniProt database (*reference*, accessed June 2022) and the derived *PAV-exclusive* and *PAV-inclusive* databases. The Proteograph™ workflow quantified 18,603 unique peptides from 2,869 proteins (**Figure 2**).

Of the 345 analyzed samples, 325 were also genotyped on the Illumina Omni 2.5 platform and had imputed genotype data available[2, 23]. Using the *PAV-inclusive* library, we identified 492 unique variant peptides that correspond to 2,341 individual signals when accounting for detections related to different nanoparticles, precursor charges, and missed cleavages (**Supplementary Table 1**). These variant peptides mapped to 317 distinct genetic variants in 251 genes. To filter to a set of reliably detected variant peptides and avoid false positives, we asked

whether each peptide's detection matched the individual blood donor's genotype. A total of 1,000 of the 2,341 variant peptide detections were significantly associated with the genotype of the coding variant in a Fisher's Exact test at a Bonferroni level of significance of $p < 2.1\times10^{-5}$ (0.05/2,341). Note that most of the non-significant variant peptides had a low detection frequency and did not provide sufficient statistical power to reach the required significance level; 512 had seven or less detections, while only peptides with eight or more detections reached Bonferroni significance. The 1,000 significant associations corresponded to 306 unique variant peptides that were generated by 184 unique MS-PAVs. These MS-PAVs were located in 137 different genes (**Supplementary Table 2**).

Robust pQTLs can be identified by excluding variant peptides from the spectral library.

In MS-proteomics, protein intensities are generally inferred from the intensities of one or multiple peptides that are derived from that protein. When peptides map ambiguously to multiple proteins, inference algorithms group them together into so-called "protein groups" and quantify them jointly. When a peptide that is used for protein quantification contains an amino acid-changing genetic variant (an MS-PAV), the resulting protein level will reflect the genotype of the sample donor and result in a spurious pQTL; this phenomenon can be considered the MS equivalent of an epitope effect in affinity proteomics. We counter this issue by excluding variant peptides from the protein quantification process. This exclusion will not necessarily reduce pQTL sensitivity for MS-PAVs that directly alter the corresponding blood protein level or for those MS-PAVs in linkage disequilibrium with regulatory variants controlling the protein's gene expression (an MS-pQTL). In these cases, all peptides derived from the protein that do not overlap with the position of the MS-PAV are expected to vary in the same way as the genotype and to equally reflect the protein level.

Figure 3 demonstrates the impact of including or excluding PAVs in the process of protein quantification in the example of the Factor V (F5) protein. When using the *PAV-exclusive* library, no pQTL is observed at the protein level, which is consistent with the absence of pQTLs on the non-PAV containing peptides. When using the *PAV-inclusive* library, the two F5 MS-PAV isoforms that correspond to a K>R substitution (rs4524) are identified as pQTLs. This is also expected, as we are considering the expression level of the isoforms separately in this case. No pQTLs are identified for any of the other peptides. However, when using the *reference*

library, a pQTL is found, which is incorrect because the overall level of F5 protein does not vary with genotype. This is an example of the MS equivalent of an epitope effect in affinity proteomics, but here in contrast to affinity proteomics, it can be easly captured.

We performed a pQTL analysis at the 184 MS-PAVs and calculated their associations with their corresponding protein intensities derived using the *PAV-exclusive* library (**Supplementary Table 2**). 14 MS-PAVs reached a genome-wide significance level of p-value < 5×10^{-8} and are considered as MS-pQTLs (**Table 1**). Half of the 14 are not found by either Olink or SOMAscan in the largest pQTL studies conducted with each; six are cross-verified by SOMAscan; and one is cross-verified by Olink. Some of these variants are found in over 40% of the study population and possess large effect sizes (absolute value > 0.5) on their respective proteins; the fact that these are missed by larger studies is of note.

We then compared the association statistics with those obtained using the *reference* library. Robust MS-pQTLs are on the diagonal of the scatterplot presented in segment 1 of **Figure 4**. Instances in which non-significant results from the *PAV-exclusive* library overlap with significant results in the *reference* library indicate situations where the current standard approach fails (segment 2 in **Figure 4**). If these variants overlap with *cis*-pQTLs from affinity proteomics studies, they reveal potential epitope effects. MS-PAVs that do not reach significance using either library are MS-PAVs that do not lead to detectable changes in protein expression in our cohort (segment 3 in **Figure 4**); however, our cohort size is relatively small and larger sample sizes are needed to reach the statistical power required to determine whether these MS-PAVs are also MS-pQTLs. These observations suggest that: a) MS-PAVs can be detected at the peptide level by using the *PAV-inclusive* library and conducting a Fisher's Exact test performed on the presence/absence of the coding variant versus MS detection/non-detection of the corresponding variant peptide; and b) using a *PAV-exclusive* library to generate pQTLs is essential to prevent false-positive pQTL identifications (i.e., the MS equivalent of epitope effects in affinity proteomics).

Overlap pQTLs from affinity proteomics platforms.

We then investigated which of the 184 MS-PAVs had been identified in previous pQTL studies (**Supplementary Table 2**). To identify overlapping pQTLs from the SOMAscan platform[10], we queried the web interface of the OmicScience server, accessed on 4 March 2023

(https://omicscience.org/apps/pgwas/pgwas.table.php). We also annotated the 184 MS-PAVs using Phenoscanner to identify pQTLs that were reported elsewhere (accessed 21 Jan 2023). To identify overlapping pQTLs from the Olink platform, we used the Supplementary Tables from the recent UKB PPP study[9]. Since this GWAS provides only the lead associations, we selected suitable proxy variants by using the variant with the highest correlation (r^2) with the MS-PAV within a 500kb window. Three additional pQTLs, not reported by these two large studies, were identified by Phenoscanner, two of which were in-*cis*.

Out of our 184 MS-PAV variants, 118 share a pQTL with at least one target of the SOMAscan platform. In 76 cases, these are *cis*-pQTLs, meaning that the SOMAscan target of a pQTL matches the protein identified by the Proteograph™ workflow. We identify 113 overlapping Olink pQTLs, 29 of which are located in-*cis*. 13 MS-PAVs are *cis*-pQTLs in both the Olink and SOMAscan study, and 92 are *cis*-pQTLs on at least one platform. 53 MS-PAVs overlap with *trans*-pQTLs in one of these studies and have no corresponding *cis*-pQTL. These MS-PAVs provide experimental evidence for the presence of the *cis*-encoded proteins via detection by mass spectrometry in blood. Of the 39 MS-PAVs with no matching pQTL in the Fenland or the UKB PPP study, three match a pQTL in another study as identified by Phenoscanner. This leaves a total of 36 MS-PAVs detected using the Proteograph™ workflow that are not previously identified in any large-scale affinity pQTL studies. These 36 novel MS-PAVs are located within 31 unique genes. Taken together, the observations highlight the complementarity between the affinity and MS-proteomics approaches.

Novel findings using the Proteograph™ workflow.

We annotated the 184 MS-PAVs with overlapping expression QTLs (eQTLs), metabolomics QTLs (mQTLs) and GWAS associations (**Supplementary Table 2**). Out of the 184 MS-PAVs, 121 match an eQTL reported in Phenoscanner, suggesting that these variants not only influence the peptide sequence but also alter the corresponding gene expression levels. 90 MS-PAVs overlap a GWAS hit (not counting metabolite/protein levels and body height), including nine of the novel variants (**Table 2**).

For example, variant rs2291725 corresponds to an S>G amino acid exchange in the Gastric Inhibitory Polypeptide (GIP). This variant is in the *GIP* gene, which codes for an incretin hormone and stimulates insulin secretion. The amino acid exchange occurs in a peptide

consisting of ten amino acids (ALELA[S/G]QANR) on the incretin pro-peptide (aa98-107). This part of the protein is not part of the processed incretin hormone. The variant peptide corresponding to the alternate allele is detected in 130 out of 223 carriers of the alternate allele, together with five false-positive detections in 102 reference allele homozygotes ($p = 1.2\times10^{-22}$, Fisher test). This variant is associated with several body fat traits, coronary artery disease, and diabetes in the GWAS catalog, consistent with *GIP's* function and suggesting a causal role for this variant in these clinical phenotypes. This *GIP* variant has not been reported by any affinity proteomics GWAS before, possibly because GIP is too small or transiently folded and may not be detected by affinity binders. GIP agonism has recently gained renewed attention as a satiety-suppressing drug (similar to GLP-1 inhibitors, but with possibly less severe side effects such as nausea)[24]. Hence, this variant may serve as a potential genetic instrument to further investigate the potential effects of GIP inhibition.

Another key protein relevant to cardiovascular disease is *APOB*. Previous GWAS studies associated genetic variation in *APOB* with many relevant lipid-related traits, with lead associations with LDL-cholesterol (LDL-C) and Apolipoprotein B (ApoB) levels measured by clinical biochemistry[25]. We found that the variant rs1367117 (chr2:21263900) associated with the alternate and also the reference allele of the ApoB variant peptide TSQC[T/I]LK (*p*-value = 1.0×10^{-68} and 3.5×10^{-16}, respectively; Fisher test), but do not detect a significant association signal at the protein level (*p*-value > 0.02). To analyze this association in its genetic context, we computed the associations of the detection of peptide TSQCILK with all variants in the vicinity (+/-250kb) of this MS-PAV, retrieved GWAS data for the associations with clinical biochemistry measures of LDL-C and ApoB in the UK Biobank, and generated regional association plots (see **Methods**, **Figure 5**). The regional association plots show that rs1367117 has the strongest LDL association, but they also indicate the presence of at least one additional equally-strong association between a variant in the promoter region of *APOB* with both LDL and ApoB levels. This observation suggests the presence of two distinct signals, one likely acting via a structural change in the ApoB protein itself, and a second that may be attributed to changes in ApoB protein levels. Interestingly, we previously found that these two signals also lead to distinct phenotypes in lipoprotein composition (see Figure S11 in Reference[26]). Our study is the first study that directly identifies this putatively causal genetic variant of high LDL-C levels at the peptide level using MS-proteomics at a population scale and shows how MS-PAVs can be used to dissect complex genetic association signals.

DISCUSSION

To the best of our knowledge, this is the first time that genetic variation has been systematically investigated at the peptide level using mass spectrometry proteomics at a population scale. We show that MS-proteomics has the potential to access genetic variation in proteins at the peptide level and to complement affinity proteomics pQTL studies by: a) providing additional information on protein identity and potential epitope effects, b) assessing proteins that are not accessible to affinity binding, and c) incentivizing future applications that elucidate post-translational modifications and protein group resolution.

Our study also has limitations. First, by excluding peptides from the *PAV-exclusive* library, some of the MS spectra remain unaccounted for and can yield false-positive matches to other peptides in the library. Future approaches could remove variant peptides only at the protein quantification step. This would also reduce the effort needed for identifying peptides with multiple libraries. In addition, such new quantification algorithms could use data from all five nanoparticles in parallel.

Another limitation is the choice of the MAF cutoff. Rarer variants are not detected at present, since including lower-frequency variants could lead to an explosion in false-positive detections. The inclusion of rare variants may also lead to multiple amino acid changes within the same peptide simultaneously, which we do not presently account for. Use of sample-specific libraries that account for individual genetic variants can mitigate this problem in the future. If these libraries additionally used phased genotype data, potential issues when two variants are located on the same peptide could also be solved.

We observe some cases where the detected variant peptide does not match the PAV, and a few isolated cases where the alternate and reference alleles are both detected in all samples, such as an E>D substitution in Complement Factor H (CFH peptide SPP[**E/D**]ISHGVVAHMSDSYQYGEEVTYK). These false-positive identifications (**Supplementary Figure X**) can be attributed to uncertainties or shortcomings in the algorithms that match the MS2 spectra of alternate and reference peptides that occur in the same DIA-MS window and share many common fragments. Due to the very low error rate in today's genotyping platforms, genetic variants can comparatively be considered a "ground truth" to

calibrate peptide detection algorithms. We suggest that these algorithms may be improved in the future by using combined genetic and proteomic data from studies like ours as a benchmark.

Taken together, our study highlights the complementarity but also the complexity of affinity- and MS-based proteomics in the pQTL discovery process and suggests a new approach to the analysis of MS-based proteomics data in the presence of genetic variation. We propose to use naturally occurring genetic variation for the development of future and more powerful MS-proteomics data analysis tools. Deployed at scale, this approach can provide valuable new insights for drug target prioritization and repurposing.

METHODS

The QMDiab study. The Qatar Metabolomics study of Diabetes (QMDiab) was conducted in 2012 at the dermatology department of Hamad Medical Corporation, the major public hospital in Doha, Qatar, with the primary aim to study metabolic differences in individuals with and without diabetes in adult female and male participants of Arab and Asian ethnicities[21, 22]. Multiple aliquots of blood, urine and saliva samples were collected and stored at -80°C without further freeze-thaw cycles.

Genotyping. DNA from QMDiab samples was extracted and genotyped using the Illumina Omni 2.5 array (version 8) and imputed using the SHAPEIT software with 1000 Genomes (phase3) haplotypes, as previously described. PAVs were identified in the imputed variant dataset using the Ensembl Variant Effect Predictor (VEP)[27] and filtering to MAF > 10%. Genotyping data was available for 325 of the 345 analyzed samples.

Gene model alignment. The gene model was constructed using the June 2022 version of the UniProt .fasta file and the UniProt genome annotation tracks (https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/genome_annotation_tracks/UP000005640_9606_beds/UP000005640_9606.proteome.bed, accessed June 2022). To create the gene model, we aligned the .bed with the .fasta they provide. We kept UniProt IDs that unambiguously mapped to one sequence. For those that mapped to multiple sequences, we preferentially selected those sequences that aligned perfectly when translating the .bed coordinates using the GRCh37 .fasta file. For those that did not map to the .fasta, we preferentially selected sequences that started with Methionine and were in-frame. We removed

ambiguity in UniProt IDs that had sequences in multiple chromosomes by picking the canonical one if available, and then alphanumerically if not. For those UniProt IDs that had multiple canonical sequences within the same chromosome, we picked the first sequence within the gene model.

Library construction. The gene model file from UniProt was used to generate reference sequences for every UniProt ID. Common (MAF > 10%) protein altering variants were identified using the Ensembl Variant Effect Predictor (VEP) [27] and "injected" into the corresponding reference protein sequences. We digested the reference and alternate sequences in a manner akin to DIA-NN *in silico* (i.e., on tryptic [K/R] amino acids; with/without one missed cleavage; and peptide length between 7-30 AAs) to generate constituent peptides per UniProt sequence. We then compared the digests from the corresponding reference and alternate gene sequences. If peptides were of equal length, shared their initial position within the full gene, and differed in sequence, then the peptides were declared a reference-to-alternate match; otherwise, they were annotated as "complex" (indicated in **Supplementary Table 1**). We repeated this process with and without missed cleavages. All other mismatched injected variant sequences (which were a result of the introduction or deletion of a K/R), were discarded. For each variant, a protein entry with the corresponding amino acid exchange was also added to the *PAV-inclusive* library as an isoform using the protein identifier (UniProt ID) followed by the variant identifier (rsID). Similarly, the corresponding reference sequences were discarded from the *PAV-exclusive* library. Genetic variants were considered independent, and only one variant per protein was considered at a time to avoid combinatorial growth of the library.

Proteomic Analysis. 240 µL of previously un-thawed citrate plasma were loaded onto the SP100 Automation Instrument for sample preparation with Proteograph™ Assay Kits and the Proteograph™ workflow[18, 19] (Seer, Inc.) to generate purified peptides for downstream LC-MS analysis. Each plasma sample was incubated with five proprietary, physicochemically-distinct nanoparticles for protein corona formation. Samples were automatically plated, including process controls, digestion control, and MPE peptide clean-up control. A one-hour incubation resulted in a reproducible protein corona around each nanoparticle surface. After incubation, nanoparticle-bound proteins were captured using magnetic isolation. A series of gentle washes removed non-specific and weakly-bound proteins. The paramagnetic property of the nanoparticles allows for retention of nanoparticles with the protein corona during each wash step. This results in a highly

specific and reproducible protein corona. Protein coronas were reduced, alkylated, and digested with Trypsin/Lys-C to generate tryptic peptides for LC-MS analysis. All steps were performed in a one-pot reaction directly on the nanoparticles. The in-solution digestion mixture was then desalted, and all detergents were removed using a solid phase extraction and positive pressure (MPE) system on the SP100 Automation Instrument. Clean peptides were eluted in a high-organic buffer into a deep-well collection plate. Equal volumes of the peptide elution were dried down in a SpeedVac (3 hours-overnight), and the resulting dried peptides were stored at -80 °C. Using the results from the peptide quantitation assay, peptides were thawed and reconstituted to a final concentration of 50 ng/µL in the Proteograph™ Assay Kit Reconstitution Buffer. 4 µL of the reconstituted peptides were loaded on an Acclaim PepMap 100 C18 (0.3 mm ID x 5 mm) trap column and then separated on a 50 cm µPAC analytical column (PharmaFluidics, Belgium) at a flow rate of 1 µL/minute using a gradient of 5 – 25% solvent B (0.1% FA, 100 % ACN) in solvent A (0.1% FA, 100% water) over 22 minutes, resulting in a 33 minute total run time. The peptides generated from these multi-nanoparticle-sampled proteins were analyzed using a dia-PASEF method[13] on a timsTOF Pro 2 mass spectrometer *(Bruker Daltonics)*.

Peptide and protein quantification. All MS files were processed using the DIA-NN 1.8.1 software[16] and a library-free search with match-between-runs (MBR) enabled against the UniProt database (accessed June 2022) and thereof derived *PAV-exclusive* and *PAV-inclusive* libraries, as described above. Peptide and protein intensities were quantified using the DIA-NN in match-between-runs mode with flags: `--mass-acc-ms1 10, --mass-acc 10, --qvalue 0.1, --matrices, --missed-cleavages 2, --met-excision, --cut K*,R*, --smart-profiling, --relaxed-prot-inf, --pg-level 1, --reannotate, --gen-spec-lib, --threads 32, --predictor, --unimod4, --use-quant, --peak-center, --no-ifs-removal,` and `--reanalyse.`

Statistical analysis. Statistical analysis was performed using R (version 4.2.1) basic libraries (fisher.test and lm) and Rstudio (version 2023.03.0). Significant MS-PAVs were identified through the construction and analysis of a 2x2 matrix. This matrix depicted how many individuals had the genetic variant at a given genomic location, and the corresponding variant peptide. We used the Fisher's Exact test to determine if there was a non-random association between these two categorical variables. Those that were statistically significant at a Bonferroni-

corrected alpha level (out of 2,341 signals) were considered for future analysis. pQTLs were determined by regressing alternate allele count to protein quantification intensities using a linear regression model. We followed previous analysis protocols[2]; i.e., we used inverse-normal scaled protein intensities as dependent variables and included age, sex, BMI, diabetes status, and the first three genetic principal components as covariates in a linear model with the copy number of the minor allele as dependent variable.

Variant annotation. The web servers snipa.org[28], phenoscanner.medschl.cam.ac.uk[29] and omicsciences.org[10] were used to identify previously reported pQTLs and overlapping information from disease GWAS, and expression and metabolomics QTLs. LocusZoom[30] was used to generate regional association plots.

APOB analysis. We downloaded UK Biobank GWAS summary statistics for LDL cholesterol (code 30780) and Apolipoprotein B (code 30640) for the joint male/female analysis with from 343,621 individuals from https://github.com/Nealelab/UK_Biobank_GWAS, extracted the +/- 250kb region around variant 2:21263900:G:A and visualized the association data using the LocusZoom server (https://my.locuszoom.org).

AUTHOR STATEMENTS

Ethics statement

The original 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. For forthgoing work with QMDiab a non-human subjects research determination was obtained.

Data availability statement

Protein and peptide level data derived using the three libraries will be publicly deposited on FigShare at time of publication.

Code availability statement

Publicly available software was used for the data analysis (Python, R and Rstudio, DIA-NN).

Acknowledgements

K.S. wishes to thank Alex Forest-Hay for initiating this project. We are grateful to all participants of QMDiab for providing their time and blood, and to the late Prof. Mohammed M. El-Din Selim for enabling the sample collection at Hamad Medical Corporation, Doha, Qatar.

Funding

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) grant NPRP11C-0115-180010. The statements made herein are solely the responsibility of the authors.

Author contributions

Study design: K.S.; Conducted Experiments: G.R.V., H.G., Data analysis: K.S., G.R.V, H.G., S.B., F.S.; Provided Materials: A.H., N.S., G.T., H.S., G.R.V., H.G.; Manuscript writing: K.S.; Manuscript editing: K.S., G.R.V., H.G., K.M., M.D., A.S., S.B., F.S. All authors contributed to the interpretation of results and critically reviewed the manuscript.

Competing interests

G.R.V., H.G., M.D., K.M., A.S., and S.B. are employees and/or stockholders of Seer, Inc.; The other authors declare no competing interests.

REFERENCES

1. Suhre K, McCarthy MI, Schwenk JM. Genetics meets proteomics: perspectives for large population-based studies. *Nature reviews Genetics* **22**, 19-37 (2021).

2. Suhre K, *et al.* Connecting genetic risk to disease end points through the human blood plasma proteome. *Nature communications* **8**, 14357 (2017).

3. Sun BB, *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73-79 (2018).

4. Sliz E, *et al.* Genome-wide association study identifies seven novel loci associating with circulating cytokines and cell adhesion molecules in Finns. *Journal of Medical Genetics* **56**, 607-616 (2019).

5. Folkersen L, *et al.* Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nature Metabolism* **2**, 1135-1148 (2020).

6. Ferkingstad E, *et al.* Large-scale integration of the plasma proteome with genetics and disease. *Nature Genetics* **53**, 1712-1721 (2021).

7. Thareja G, *et al.* Differences and commonalities in the genetic architecture of protein quantitative trait loci in European and Arab populations. *Human Molecular Genetics*, (2022).

8. Dhindsa RS, *et al.* Influences of rare protein-coding genetic variants on the human plasma proteome in 50,829 UK Biobank participants. *bioRxiv*, 2022.2010.2009.511476 (2022).

9. Sun BB, *et al.* Genetic regulation of the human plasma proteome in 54,306 UK Biobank participants. *bioRxiv*, 2022.2006.2017.496443 (2022).

10. Pietzner M, *et al.* Mapping the proteo-genomic convergence of human diseases. *Science (New York, NY)* **374**, eabj1541 (2021).

11. Pietzner M, *et al.* Synergistic insights into human health from aptamer- and antibody-based proteomic profiling. *Nature communications* **12**, 6822 (2021).

12. Enroth S, Johansson A, Enroth SB, Gyllensten U. Strong effects of genetic and lifestyle factors on biomarker variation and use of personalized cutoffs. *Nature communications* **5**, 4684 (2014).

13. Meier F, *et al.* diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition. *Nature Methods* **17**, 1229-1236 (2020).

14. Schessner JP, Voytik E, Bludau I. A practical guide to interpreting and generating bottom-up proteomics data visualizations. *Proteomics* **22**, e2100103 (2022).

15. Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature* **422**, 198-207 (2003).

16. Demichev V, *et al.* dia-PASEF data analysis using FragPipe and DIA-NN for deep proteomics of low sample amounts. *Nature communications* **13**, 3944 (2022).

17. Peckner R, *et al.* Specter: linear deconvolution for targeted analysis of data-independent acquisition mass spectrometry proteomics. *Nat Methods* **15**, 371-378 (2018).

18. Liu Y, Wang J, Xiong Q, Hornburg D, Tao W, Farokhzad OC. Nano-Bio Interactions in Cancer: From Therapeutics Delivery to Early Detection. *Accounts of chemical research* **54**, 291-301 (2021).

19. Ferdosi S*, et al.* 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 (2022).

20. Blume JE*, et al.* Rapid, deep and precise profiling of the plasma proteome with multi-nanoparticle protein corona. *Nature communications* **11**, 3662 (2020).

21. Mook-Kanamori DO*, et al.* 1,5-Anhydroglucitol in saliva is a noninvasive marker of short-term glycemic control. *The Journal of clinical endocrinology and metabolism* **99**, E479-483 (2014).

22. Yousri NA*, et al.* 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 (2015).

23. Sharapov SZ*, et al.* Defining the genetic control of human blood plasma N-glycome using genome-wide association study. *Hum Mol Genet* **28**, 2062-2077 (2019).

24. Hayes MR, Borner T, De Jonghe BC. The Role of GIP in the Regulation of GLP-1 Satiety and Nausea. *Diabetes* **70**, 1956-1961 (2021).

25. Sinnott-Armstrong N*, et al.* Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat Genet* **53**, 185-194 (2021).

26. Suhre K*, et al.* Lipoprotein profile and metabolic fine-mapping of genetic lipid risk loci. *medRxiv*, 2022.2006.2012.22276286 (2022).

27. McLaren W*, et al.* The Ensembl Variant Effect Predictor. *Genome Biology* **17**, 122 (2016).

28. Arnold M, Raffler J, Pfeufer A, Suhre K, Kastenmüller G. SNiPA: an interactive, genetic variant-centered annotation browser. *Bioinformatics (Oxford, England)* **31**, 1334-1336 (2015).

29. Kamat MA*, et al.* PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. *Bioinformatics (Oxford, England)* **35**, 4851-4853 (2019).

30. Boughton AP*, et al.* LocusZoom.js: interactive and embeddable visualization of genetic association study results. *Bioinformatics (Oxford, England)* **37**, 3017-3018 (2021).

TABLES

Table 1. MS-pQTLs. Associations of MS-PAVs that are significantly ($p < 5\times10^{-8}$) associated with protein levels derived using the *PAV-exclusive* library.

Gene	UniProtID	rsid	SNP	MAF	*p*-value	beta	*cis*-pQTL
MST1	G3XAK1	rs3197999	3:49721532:G:A	23.4%	2.6E-18	-0.716	SOMA
ITIH1	P19827	rs1042779	3:52821011:A:G	37.7%	1.2E-11	0.545	SOMA
KNG1	P01042	rs2304456	3:186445052:T:G	16.0%	4.8E-21	0.911	SOMA
HLA-C	A2AEA2	rs707908	6:31238053:G:C	21.8%	2.0E-10	0.630	no
CFB	B4E1Z4	rs12614	6:31914179:C:T	18.5%	1.2E-13	-0.590	SOMA
PON1	P27169	rs662[+]	7:94937446:T:C	37.4%	4.6E-09	-0.424	no
PON1	P27169	rs854560[+]	7:94946084:A:T	27.4%	2.7E-10	0.499	no
PON2	A0A0J9YXF2	rs12026	7:95041016:G:C	30.6%	1.3E-36	-0.998	OLINK
FGL1	Q08830	rs3739406	8:17739538:T:C	49.2%	1.8E-11	0.512	SOMA
GALC	G3V255	rs34362748	14:88442712:C:T	11.5%	1.0E-15	0.887	no
SERPINA10	G3V2W1	rs2232700	14:94756450:T:A	30.8%	1.5E-22	0.741	SOMA
SERPINA1	P01009	rs709932	14:94849201:C:T	23.1%	3.8E-08	0.431	no
DSC3	Q14574	rs276938[*]	18:28610988:C:T	41.4%	2.1E-08	-0.422	no
DSC3	Q14574	rs276937[*]	18:28611061:A:T	41.2%	1.5E-08	-0.429	no

[+]correlation between rs662 and rs854560 is $r^2 = 0.20$
[*]correlation between rs276938 and rs276937 is $r^2 = 0.99$

Table 2. MS-PAVs that overlap with disease-relevant GWAS hits. Selected MS-PAVs that have not been reported in previous pQTL studies and that overlap with a clinically relevant GWAS catalog entry (edited for brevity, see **Supplementary Table 2** for details).

Gene	UniProtID	rsID	Fisher P	Peptide	GWAS trait
WDR1	O75083	rs13441	4.8E-37	FTIGDHSR	Atrial fibrillation and flutter
PIP4K2A	H7BXS3	rs2230469	2.2E-48	IYIDDNSK	Body fat composition
CHID1	Q9BWS9	rs6682	5.9E-59	MVWDSQASEHFFEYK	Body mass index
SCFD1	A0A7I2V362	rs229150	1.8E-28	FGQDIISPLLSVK	Amyotrophic lateral sclerosis
SPTB	P11277	rs229587	9.6E-14	ETWLNENQR	Red blood cell phenotypes
LOXL1	H3BUV8	rs1048661[+]	1.0E-70	EVAVGDSTGMALAR	Exfoliation glaucoma
LOXL1	H3BUV8	rs3825942[+]	5.5E-57	HGDSASSVSASAFASTYR	Coronary artery disease, Exfoliation glaucoma
ACAN	A0A5K1VW97	rs1126823	3.9E-25	ITCTDPTTYK	Osteoarthritis
GIP	P09681	rs2291725	1.2E-22	ALELAGQANR	Coronary artery disease, Type II diabetes

[+]correlation between rs1048661 and rs3825942 is $r^2 = 0.12$

FIGURES

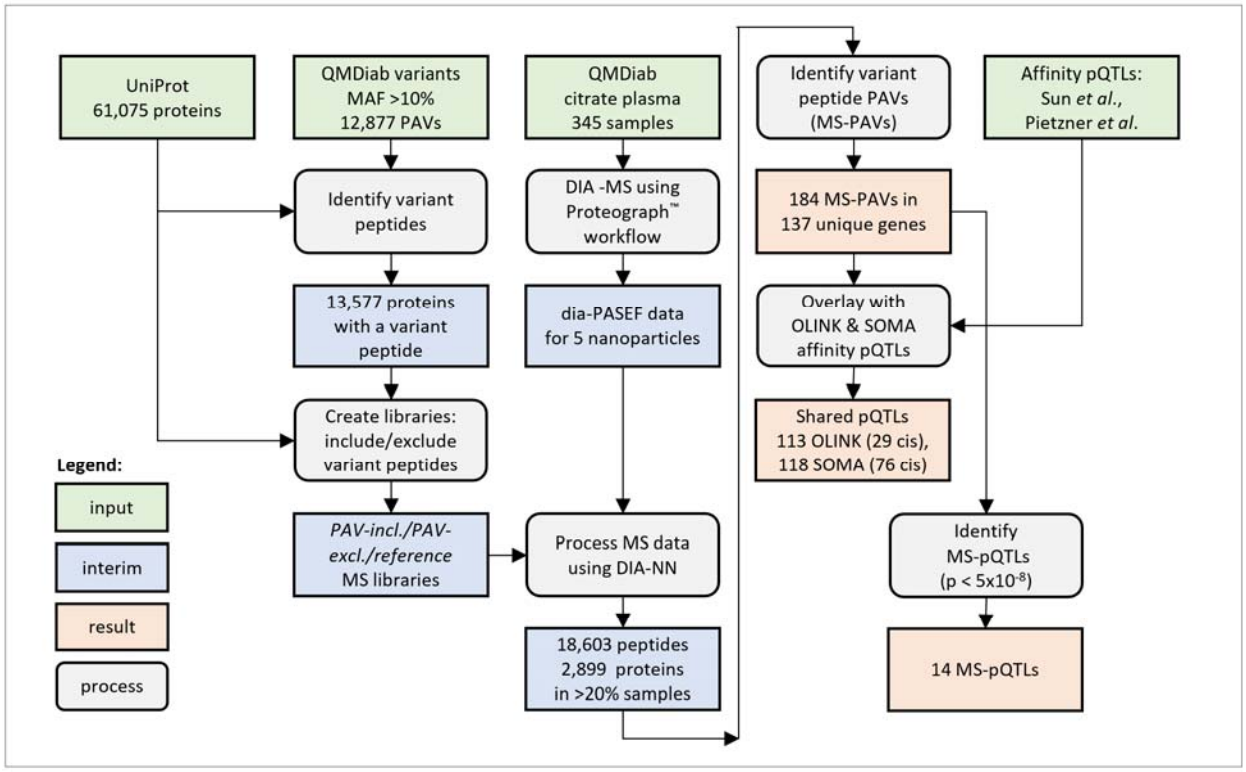
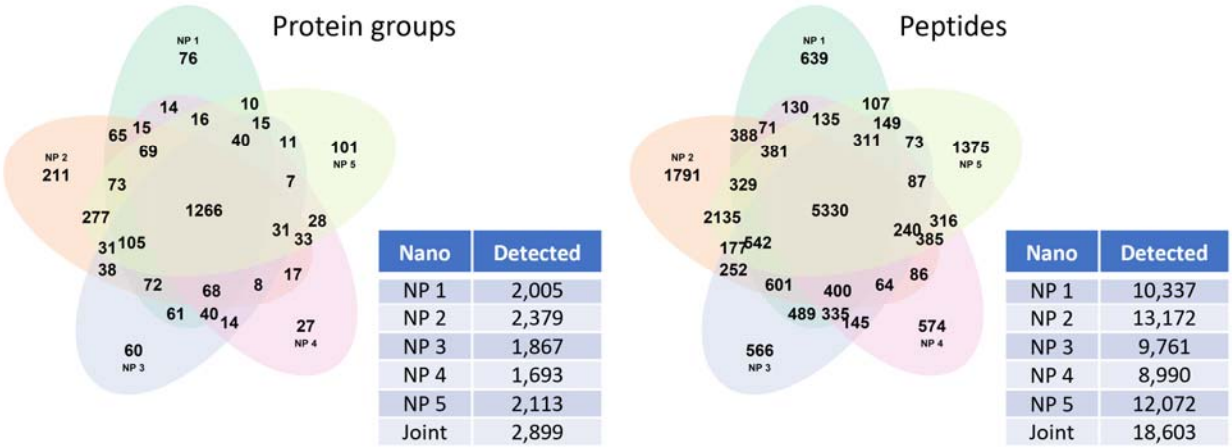


Figure 1: Study chart. Procedure used to incorporate QMDiab variants into the UniProt .fasta, create spectral libraries, and identify MS-PAVs and MS-pQTLs.



Nano	Detected
NP 1	2,005
NP 2	2,379
NP 3	1,867
NP 4	1,693
NP 5	2,113
Joint	2,899

Nano	Detected
NP 1	10,337
NP 2	13,172
NP 3	9,761
NP 4	8,990
NP 5	12,072
Joint	18,603

Figure 2. Proteins and peptides identified in >20% of the samples by the Proteograph™ workflow. Data is for protein and peptide identification using DIA-NN with the reference (*ref*) library using match-between-runs (MBR).

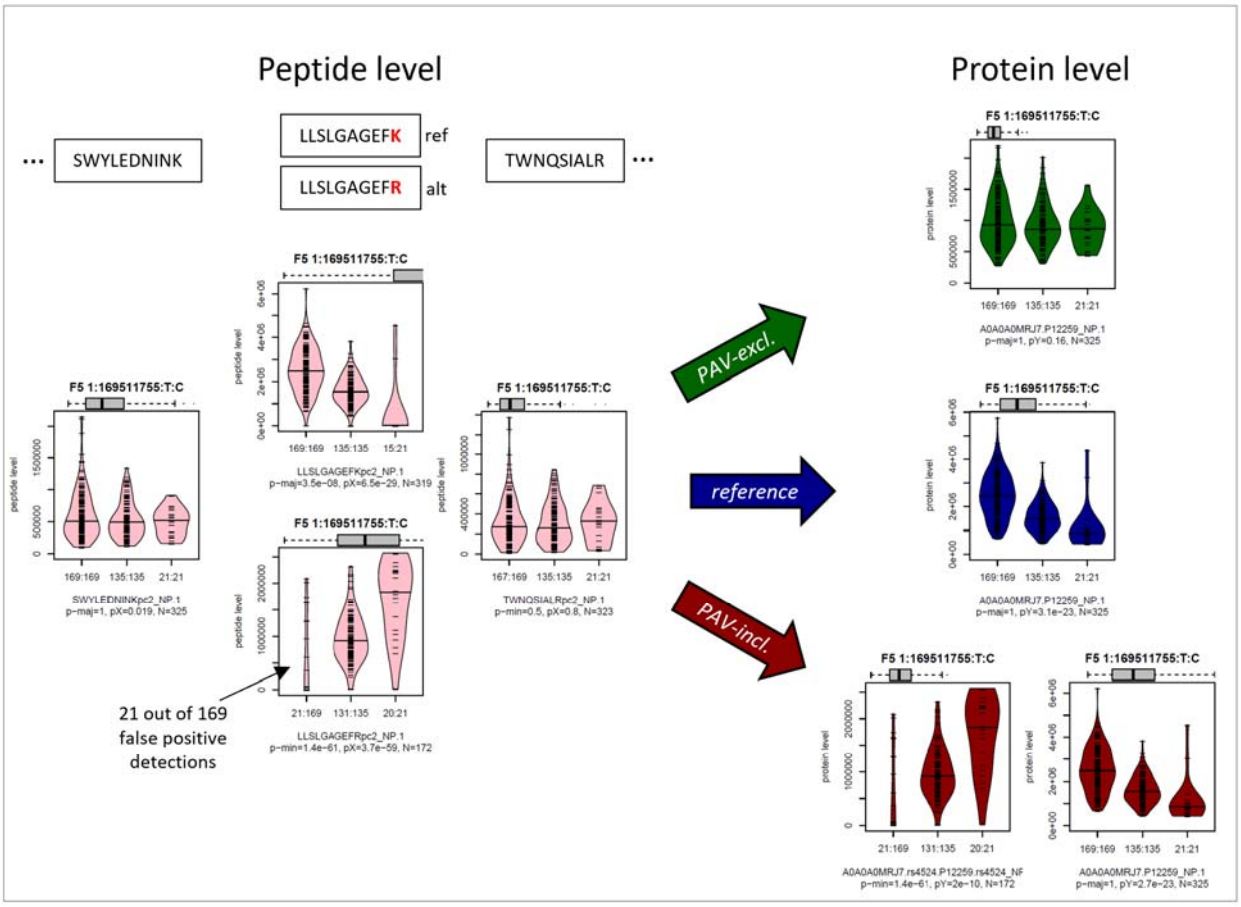


Figure 3. Boxplots by genotype rs4524 for selected Factor V (F5) protein and peptide intensities. This figure shows the effect of using the different libraries at the example of the Factor V (F5) protein. Similar plots are provided as **Supplementary Figure 1** for all 184 MS-PAVs; The boxes are color-coded as following: using the *PAV-exclusive* library (green), using the *reference* library (blue), and using the *PAV-inclusive* library (red). Protein intensities are in dark colors, and peptide intensities are in light colors. The grey vertical boxplots on top of the plots represent the range of the data shown in that plot compared to the 5%-95% range of the entire data for that protein. Units on the *y*-axis are engine-normalized intensities as provided by DIA-NN. The *x*-axis labels indicate the number of detected peptides/proteins followed by a colon and the number of samples with the given genotype (order: reference/major allele homozygote, heterozygote, alternate/minor allele homozygote). The first line of the subtitle identifies the protein (Uniprot ID and rsID, when applicable) or the peptide sequence followed by the nanoparticle used in that analysis. The second line shows the number of data points included in generating the plot (*N*). Significance intensities (*p*-values) for the following hypothesis tests are given: (1) Fisher's Exact test on detected/non-detected versus presence/absence of the major (p-maj) or minor (p-min) allele, where the stronger of the two associations is shown (indicating MS-PAV detection significance), and (2) a linear regression of peptide intensity versus genotype (coded 0-1-2) with missing values set to zero (pX), and for proteins a linear model including relevant covariates using inverse-normal scaled protein intensities (excluding missing values) against genotype (pY; indicating pQTL significance). Protein name, chromosome, chromosome position (GRCh37), and major and minor alleles are indicated in boldface on top of the boxplots.

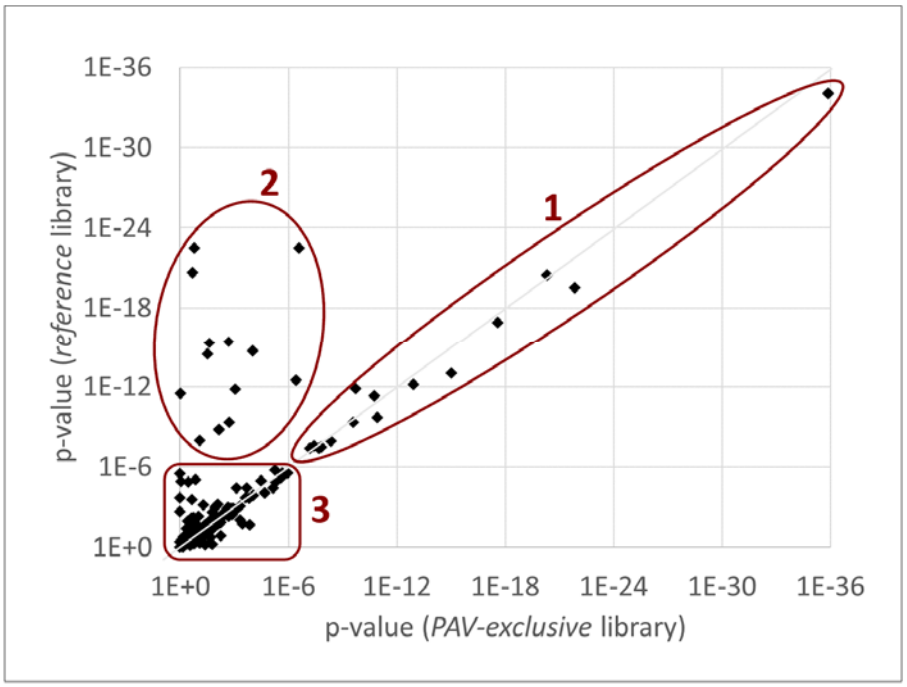


Figure 4. Scatterplot of the protein-level associations (*p*-values) for the 184 MS-PAVs using the *reference* and the *PAV-exclusive* libraries. Three regimens are labeled: **(1)** variants that remain associated with protein levels after removal of the variant peptides from the library (MS-pQTLs), **(2)** variants where the association signal with the protein levels disappears after removal of the variant peptides (the MS equivalent of an epitope effect), and **(3)** variants that do not associate with protein levels in either case (MS-PAVs that may become significant in more highly powered studies). Plot data is in **Supplementary Table 2**.

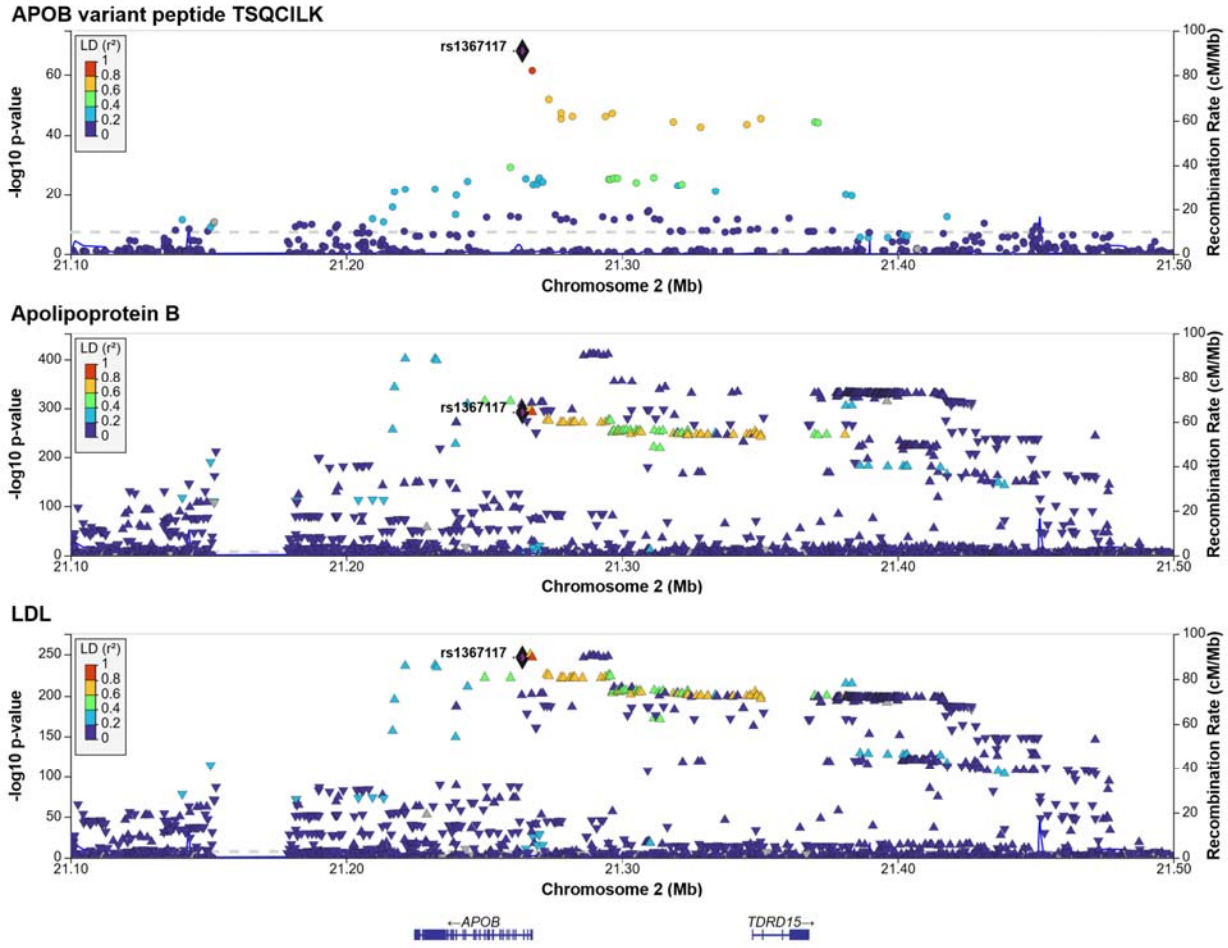


Figure 5: Regional association plots for the APOB region. Association of the detection of the alternate variant peptide TSQCILK of APOB (pc2, nanoparticle 1) with the presence/absence of the matching genetic variants at the APOB locus (top), GWAS associations of Apolipoprotein B (middle) and LDL-cholesterol (bottom) measured by clinical biochemistry methods in blood samples from 343,621 participants of the UK Biobank study. The highlighted variant rs1367117 (chr2:21263900) is the MS-PAV in TSQC[T/I]LK.

SUPPLEMENTARY TABLES

Supplementary tables are provided in EXCEL format.

ST1	List of all detected variant-peptides (2,341 detections, often on multiple nano particles and with varying precursor charges, for a total of 492 unique peptides)
ST2	List of significant lead variant to variant-peptide associations (N = 184)