Title

Peptide-centric analyses of human plasma enable increased resolution of biological insights into non-small cell lung cancer relative to protein-centric analysis

Authors

Margaret K. R. Donovan[1,†,*], Yingxiang Huang[1,†], John E. Blume[1], Jian Wang[1], Daniel Hornburg[1], Iman Mohtashemi[1], Sangtae Kim[1], Marwin Ko[1], Ryan W. Benz[1], Theodore L. Platt[1], Serafim Batzoglou[1], Omid C. Farokhzad[1,*], and Asim Siddiqui[1,*]

[1]Seer, Inc., Redwood City, CA, USA

[†]These authors contributed equally.
[*]Correspondence to: Margaret K.R Donovan, Omid C. Farokhzad, and Asim Siddiqui

Abstract

Comprehensive assessment of the human proteome remains challenging due to multiple forms of a protein, or proteoforms, arising from alternative splicing, allelic variation, and protein modifications. As proteoforms can serve distinct functions and act as functional links between genotype and phenotype, proteoform-level knowledge is critical in understanding the molecular mechanisms underlying health and disease. However, identification of proteoforms requires unbiased protein coverage at amino acid resolution. Scalable, deep, and unbiased proteomics studies have been impractical due to cumbersome and lengthy workflows required for complex samples, like blood plasma. Here, we demonstrate the power of the Proteograph™ Product Suite in enabling unbiased, deep, and rapid proteomics at scale in a proof-of-concept proteoform analysis to dissect differences between protein isoforms in plasma samples from 80 healthy controls and 61 patients with early-stage non-small-cell lung cancer (NSCLC). Processing the 141 plasma samples with Proteograph yielded 22,993 peptides corresponding to 2,569 protein groups at a confidence of 1% false discovery rate. We extracted four proteins with peptides with significant abundance differences ($p < 0.05$; Benjamini-Hochberg corrected) in healthy control and cancer plasma samples. For one, the abundance variation can be explained by underlying annotated protein isoforms. For a second, we find evidence for differentially transcribed isoforms in the broader sequence data, but not in the known annotated protein isoforms. The others may be explained by novel isoforms or post-translational modifications. In addition, we sought to identify protein variants arising from allelic variation. To this end, we first performed whole exome sequencing on buffy coat samples from 29 individuals in the NSCLC study. Then, we created personalized mass spectrometry search databases for each individual subject from the exome sequences. From these libraries, we identified 422 protein variants, one of which has previously been shown to relate to lung cancer. In conclusion, our results demonstrate that Proteograph can generate unbiased and deep plasma proteome profiles that enable identification of proteoforms present in plasma at a scale sufficient to enable population-scale proteomic studies powered to reveal novel mechanistic and biomedical insights.

Introduction

Multiple isoforms of a single protein, or proteoforms, can arise due to alternative splicing (i.e., protein isoforms), allelic variation (i.e., protein variants), and post translational modifications (PTMs) [1]. Proteoforms play key and distinct roles in biological mechanisms, including impacting complex traits [2] and disease [3–5]. Genetic variation gives rise to changes to the genome that can be functionally neutral, however some variants, such as non-synonymous variants resulting in the alteration of an amino acid sequence (i.e., protein variants), can drastically impact phenotype [6]. Importantly, rare variants are highly enriched for pathogenicity (i.e., are much more likely to be deleterious and to have a large effect in common and rare disease) and common variants are known to usually be either benign, or to have a small effect in disease [7–9]. In a population, rare genetic variants substantially outnumber common variants [10]. Comprehensive examination of proteoforms in a complex sample remains elusive. Each proteome harbors a large fraction of putatively physiologically relevant rare proteoforms. These cannot easily be accessed with protein affinity-based targeted methods, particularly because there are estimated to be over 1 million distinct proteoforms in a given cell type [11]. Thus, targeted methodologies designing a panel comprising all these potential proteoforms is impractical. We show untargeted approaches that identify both protein isoforms and protein variants provide a deeper, more nuanced assessment of human proteomes, supporting enhanced understanding of human health and disease.

Important advances in characterizing the proteomic landscape of lung cancers such as non-small cell lung cancer (NSCLC) and squamous cell lung cancer have identified important protein biomarkers [12–14]. However, few proteoforms relevant to lung cancer have been identified [15]. Unbiased readout technologies such as high-resolution quantitative mass spectrometry (MS) can be employed to infer and quantify peptides and proteins with high confidence (e.g., < 1% false discovery rate (FDR)). However, large-scale LC-MS/MS-based proteomics studies have largely been impractical due to cumbersome and lengthy workflows required to achieve deep (i.e., broad detection of proteins across the dynamic range, from high to low abundance proteins) and unbiased (i.e., hypothesis-free detection) sampling of clinically relevant biospecimens with large dynamics ranges of protein abundances, such as blood plasma [16–18]. Further, while LC-MS/MS methodologies have the technological capability to infer proteoforms, peptide identification in LC-MS/MS-based proteomic data relies on protein databases, such as UniProt [19], which exclude most proteoforms that may be present in an individual's proteome.

Recently, the Proteograph Product Suite, a nanoparticle (NP)-based sampling methodology upstream of LC-MS/MS-based workflows, has demonstrated its ability to enable fast, scalable, deep, and unbiased plasma proteomics [20]. While Proteograph was able to identify both known and novel NSCLC biomarker sets, its ability to detect NSCLC-relevant protein variants has not yet been demonstrated. An important advantage to untargeted LC-MS/MS-based proteomic data is its ability to be reinterrogated with new biological questions and computational advances. To infer proteoforms arising from alternative splicing, it is possible to use Proteograph-identified peptides and abundances to observe examples of alternative exon usage. Further, to identify proteoforms arising from genetic variation, it is possible to use subject-matched genotype data, such as whole exome sequencing (WES) data, to generate custom protein databases. Both strategies offer the potential to identify proteoforms that otherwise would not be identified using protein affinity-based targeted technologies.

Here, we analyzed the proteomes from a cohort of healthy individuals, individuals with early and late NSCLC, and individuals with comorbidities using Proteograph to explore our ability to infer proteoforms. In data independent acquisition (DIA) data generated from 141 subjects (80 healthy subjects and 61 subjects identified as having early NSCLC), we used a discordant peptide intensity search (Figure 1A) to infer four proteins with differentially abundant protein isoforms, including BMP1, which is known to play both an activator and repressor role in cancer [21]. Further, in data dependent acquisition (DDA) data generated from 29 subjects (11 healthy subjects, 5 early NSCLC subjects, 9 late NSCLC subjects, and 4 subjects with comorbidities), we used a proteogenomic search (Figure 1A), which incorporated WES data to generate personalized databases, and we identified 422 proteins variants. Together, these findings demonstrate Proteograph-based proteomic data can be used to identify proteoforms and emphasizes proteoform inference is enabled by LC-MS/MS-based, unbiased approaches.

Results

Peptide-level analyses provides unique biological insight at the peptide-level versus protein-level

To assess the ability to detect proteoforms (protein isoforms and protein variants) in LC-MS/MS-based plasma proteomic data derived from Proteograph, we obtained DIA data generated from Proteograph performed on 141 subjects (80 healthy subjects and 61 subjects identified as having early NSCLC, hereto referred to as "early NSCLC subjects") using physiochemically distinct 10 physiochemically distinct nanoparticles (NP) (Figure 1A) [20]. Data was analyzed using Spectronaut (Biognosys, Switzerland) for peptide identification and protein group assembly, as described previously [20]. Across the 141 samples, we detected 2,569 unique protein groups (hereto referred to as *proteins*) and 2,010 proteins identified in at least 25% of subjects (Figure 1B). Mapping to the 2,569 unique proteins, we identified 22,993 peptides, with a median of 5 peptides per protein and mean of 8.8 peptides per protein (Figure 1C). To examine the extent to which these proteins cover the broad dynamic range of the human plasma proteome, we ranked the 2,569 proteins using abundances derived from the Human Plasma Proteome Project (HPPP) [22] (Figure 1D). Ranking our proteins by the corresponding protein abundances in the HPPP reference from lowest to highest, we observed that the first quartile, median, and third quartile are 0.785, 8.8, and 190 ng/ml, respectively. On the other hand, when proteins found in a depleted plasma approach was ranked by abundance in the HPPP reference, we observe that the first quartile, median, and third quartile are 27, 190, and 2,600 ng/ml, respectively. These results demonstrate that Proteograph can extract lower abundant proteins. Further, these results indicate Proteograph-derived proteins have suitable depth and diversity to perform proteoform discovery analyses.

To examine if abundance differences between healthy and early NSCLC subjects were detected, we searched for proteins and peptides that are differentially abundant (DA). First, to reduce potential noise introduced by rare peptides, proteins were filtered to those present in at least 50% of subjects from either heathy or early NSCLC, retaining 10,280 peptides and 1,565 proteins across 141 subjects (Figure 1E). Next, as each protein may have been detected by more than one NP, we use MaxLFQ [23] to quantify a single abundance (hereto referred to as *collapsed abundances*) between healthy and early NSCLC subjects. We evaluated differential protein abundances observing 243 significantly regulated proteins (adjusted $p < 0.05$; Wilcoxon Test) (Figure 1F, G). To investigate NPs capacity to capture biological signal beyond abundance levels (e.g., proteoform information, or NP specific protein complexes), we treated each NP:protein feature pair as a separate observation comparing healthy and early NSCLC subjects. Here, we identified 877 NP:protein feature pairs (Figure 2F), corresponding to a 3.6-fold increase from examining differences at the aggregated level alone. This highlights the capacity of NPs to interrogate the proteome: the signal they capture can be more biologically relevant than that captured by conventional DA analysis. Finally, we performed differential abundance analysis using peptide abundances across all NPs (i.e., not collapsed abundances) between healthy and early NSCLC subjects and identified 1,581 DA peptides (Figure 1F, H), corresponding to a 6.5-fold increase from examining differences at the protein-level. These results indicate that by leveraging both NP specific features patterns and peptide information, we can increase the number of observable differences, which may lead to a better classification of phenotypes and additional mechanistic insights.

To test the utility of NP and peptide-level interrogation of complex biological samples, we examined the top 10 most DA proteins and peptides. Among the top DA peptides, we observed peptides mapping to ITIH2, ANTXR2, and ANTXR1, which are known to be downregulated in early NSCLC plasma samples. Downregulation of ITIH2 expression has been seen in 70% of breast cancers, 71% of lung cancers, and 70% of renal tumors [24]. ANTXR2/CMG2 was shown to inhibit breast cancer cell growth and is inversely correlated with disease progression and prognosis [25]. ANTXR1 can reduce tumor growth in vivo by targeting cancer stem cells in conjunction with LeTx [26]. In agreement with results of other studies, our analysis of NSCLC plasma samples showed upregulation of well-defined pro-inflammatory and cancer biomarkers such as CRP, S100A9, and S100A8 [27,28]. Together, the observation of known hallmark cancer and inflammatory biomarkers indicates Proteograph-derived proteomic data captures known biological differences and may suggest the presence of other novel biomarkers. Overall, this increased number of observed significant differences between proteins, protein across NPs, and peptides across NPs, verified by the presence of known cancer biomarkers, indicates substantial opportunity to increase biological insight and potential to identify proteoforms by increasing the resolution used examine our proteomic data.

Identification of four NSCLC-associated proteoforms using peptide-level discordant peptide search

We next explored whether we could use DA peptides in contrast to the average protein-level information to help resolve proteoforms. Specifically, we extracted DA peptides and retained proteins with at least one peptide over-expressed in healthy subjects and at least one peptide over-expressed in early NSCLC subjects (Figure 2A). Then, by mapping the DA peptides to genomic space, we could infer potential exon usage and proteoforms. We performed this discordant peptide intensity analysis and identified four proteins for which we potentially captured multiple protein isoforms with significant differential behavior in early NSCLC when compared to healthy controls: BMP1, C4A, C1R, and LDHB (Figure 2B). We examined the Open Target Score [29], which is an association score of known and potential drug targets with diseases using integrated genome-wide data from a broad range of data sources, to assess the association of the four proteins with lung carcinoma targets. We found modest to low scores (Figure 2B), suggesting a mix of novel and known lung cancer-relevant proteins. These proteins have all been previously identified in plasma and range from highly abundant (C4A, C1R, LDHB) to moderately abundant (BMP1) [22] (Figure 2C). BMP1, the least abundant of the four proteins, is not identified in depleted plasma, indicating this approach identified protein isoforms inaccessible with conventional depleted plasma proteomics workflows. These results indicate that, using a MS-based peptide discordant intensity search, we identify proteoforms with possible relevance to NSCLC.

To interrogate the different potential isoforms of these proteoforms, we examined differences in abundances between healthy and early NSCLC subjects for each of the four at the collapsed protein level, NP:protein level, and peptide level. First examining BMP1, at the collapsed protein (Figure 2D) and NP:protein (Figure 2E) level, we do not observe a difference in BMP1 abundance, likely due to averaging of peptide abundances occurring at the protein-level. However, at the peptide-level (Figure 2F), there are three significantly differential peptides: 1) peptide 1, which is significantly upregulated in early NSCLC subjects (adjusted $p = 5.29 \times 10^{-4}$; Wilcoxon Test); 2) peptide 3, which is significantly upregulated in healthy subjects (adjusted $p = 1.21 \times 10^{-2}$; Wilcoxon Test); and 3) peptide 7, which is significantly

upregulated in healthy subjects (adjusted $p = 4.99 \times 10^{-2}$; Wilcoxon Test). We also observe a trend in direction of abundance differences, where the first two peptides are upregulated in early NSCLC subjects and the last five peptides are upregulated in healthy subjects (Figure 2F). To assess whether these two groups of peptides belong to different proteoforms, we further compared their abundance similarities across the 141 subjects. We expect peptides that belong the same proteoform should have correlated abundances across individuals since they belong to the same molecular entity while peptides belong to different proteoforms should have non-correlated abundances across a cohort of individuals. We thus performed pairwise Pearson correlation and hierarchical clustering analysis, which showed two distinct clusters driven by a high degree of correlation in peptide 1 and 2 (cluster 1) and peptides 3-7 (cluster 2) (Figure 2G). We next mapped the peptides to their genomic sequence, including four protein coding isoform transcripts (ENST00000397814, ENST00000354870, ENST00000306349, and ENST00000306385), and ordered them according to exon order (Figure 2H). We observed two distinct segments of corresponding direction of BMP1 peptide differential abundance. Specifically, peptides 1 and 2 were both upregulated in early NSCLC subjects (segment 1) and peptide 3-7 were all upregulated in healthy subjects (segment 2) (Figure 2F). Peptides mapping segment 1 exclusively map to exons present in the short isoform (ENST00000397814), whereas peptides mapping to segment 2 exclusively map to exons present in the three longer isoforms (ENST00000354870, ENST00000306349, and ENST00000306385) (Figure 2H). Since segment 1 peptides, corresponding to the short BMP1 isoform, are upregulated in early NSCLC subjects and segment 2 peptides, corresponding to the longer BMP1 isoforms are upregulated in healthy subjects, these results may suggest distinct consequences of the long and short BMP1 isoforms. As BMP1 is known to play a dual role in cancer, acting as both a suppressor and activator [21], these results may reveal a role of BMP1 isoform abundance in cancer.

To interrogate the potential different isoforms, we next examined C4A. At the collapsed protein (Supplemental Figure 1A) and NP:protein (Supplemental Figure 1B) level, as for BMP1, the difference in C4A abundance is not statistically significant, however at the peptide level there are many significantly differentially expressed peptides (Supplemental Figure 1D). Like BMP1, this further supports protein abundance comparisons mask differences that occur at the peptide level. Pairwise Pearson correlation and hierarchical clustering analysis showed two distinct clusters driven by peptide abundance correlation in peptides 40-54 (cluster 1) and peptides 1-39 and 55-64 (cluster 2) (Supplemental Figure 1C). Mapping the peptides to the two known protein coding isoform transcripts (ENST00000428956 and ENST00000498271) and ordering them according to exon order (Supplemental Figure 1E), we observed three distinct segments of corresponding direction of C4A peptide differential expression abundance. Specifically, peptides 1-39 were upregulated in healthy subjects (segment 1), peptides 40-54 were upregulated in early NSCLC subjects (segment 2), and peptides 55-64 were upregulated in healthy subjects (segment 3) (Supplemental Figure 1D). Interestingly, segments 1 and 3 correspond to cluster 2 and segment 2 corresponds to cluster 1, indicating discordant expression of NSCLC-associated exons in segment 2. In contrast to results from analysis BMP1, there was no obvious association with the two known C4A protein coding isoform transcripts. However, he discordant peptides of segment 2 suggest the presence of an unknown isoform or a smaller byproduct from C4A or other members of the C4 complex. [30]

To interrogate the other potential isoforms, we next examined C1R and LDHB. At the collapsed protein (Supplemental Figure 2A, 3A) and NP:protein (Supplemental Figure 2B, 3B) level, the difference in C1R and LDHB abundance, respectively, was not statistically significant. However, at the peptide level we observe differentially expressed peptides (Supplemental Figure 2D, 3D), indicating protein abundance comparisons mask differences that occur at the peptide level. Pairwise Pearson correlation and hierarchical clustering analysis showed one distinct cluster in each protein. For C1R, this consisted of moderately correlated peptides 1, 2, 4, 6-11, 14, 16, and 17 (cluster 1) and a weak correlation between peptides 3, 5, 12, 13, and 15 (Supplemental Figure 2C). For LDHB, this consisted of highly correlated peptides 1-3, 6, and 8-11 (cluster 1) and a weak correlation between peptides 4, 5, and 7 (Supplemental Figure 3C). We mapped the peptides to the known protein coding isoform transcripts (C1R: ENST00000647956, ENST00000536053, ENST00000535233, ENST00000649804, ENST00000543835, and ENST00000540242; LDHB: ENST00000647956, ENST00000536053, ENST00000535233, ENST00000649804, ENST00000543835, and ENST00000540242) and ordered them according to exon order (Supplemental Figure 2E, 3E). There was no clear pattern in healthy or NSCLC subject peptide upregulation corresponding to any of the known isoforms for either protein. However, we did observe upregulation in C1R peptides 14-17 in healthy subjects corresponding to the two short isoforms (isoforms 5 and 6; Supplemental Figure 2E). Beyond this observation, we could not explain the discordance in peptide abundances Further examination of C1R and LDHB protein isoforms is needed to further explain the discordance in peptide abundances. Together, these results indicate discordant peptide abundance can be utilized to identify some disease-relevant protein isoforms, as was observed in the case of BMP1 and C4A. However, this approach cannot explain all discordance, as was observed in the case of C1R and LDHB. Further work may expand our understanding of disease-associated proteoforms.

Proteoforms arising from genetic variation can be identified using a proteogenomic approach

We obtained DDA data from 29 subjects (11 healthy subjects, 5 early NSCLC, 9 late NSCLC subjects, and 4 comorbid subjects), for which WES data was generated. This data was utilized to perform custom proteogenomic searches (i.e., incorporating WES-derived variant information to generate personalized databases) and to identify protein variants (Figure 5A). Specifically, for each subject we identified single nucleotide variants (SNVs) that result in single amino acid variants (SAAVs) and used to generate custom peptide sequences that could exist in each individual subject's proteome (i.e., personalized peptide sequences). Using these personalized databases, we searched for protein variants and across all 29 subjects and mapped 422 protein variants with an average of 79.59 ± 23.57 protein variants per subject (Figure 5B). We examined the alternative allele frequencies of identified protein variants and found the distribution to follow that seen in population-scale studies [31], including the observation of rare alleles (Figure 5C). For example, we detected a peptide variant harboring a SAAV (H→R) resulting from previously identified genetic variant related to lung cancer, rs1229984 [32], in one early NSCLC individual. This genetic variant changes a histidine to arginine, leading to a detectable short alternative variant peptide. Interestingly, we did not observe the reference peptide likely due to the long segment of the protein without an arginine or lysine site for trypsinization. Additionally, we observed peptides in which we identified both the reference and

alternative alleles from heterozygous variants. For example, in an early NSCLC subject, 021-004, we detect a SAAV (N→I), for which we identify both the reference (HPLKPDNQPFPQSVSESCPGK) and alternative peptide (HPLKPDIQPFPQSVSESCPGK) mapping to HRG (Figure 5D), indicating we recapitulate peptide search results from standard searches in addition to capturing novel and personalized peptides. Together, these results demonstrate the potential of peptide-centric proteogenomic and NP-based workflow for gaining new biological insights into NSCLC.

Discussion

Existing technologies, including an unbiased NP-based methodology upstream LC-MS/MS-based workflows and targeted methodologies, have enabled protein-centric analyses that have revealed important insights into human disease. While protein-centric analyses have made substantial strides in our understanding of human biology, protein-level studies may conceal biologically critical features, like proteoforms arising from alternative splicing (protein isoform), allelic variation (protein variants), or post-translational modifications, which provide mechanistic insights underlying complex traits and disease. Importantly, unbiased LC-MS/MS-based proteomic data can be re-mined and enable peptide-centric analyses that may reveal new information about proteoforms. The rationale for this study is that peptide-level information can be derived from LC-MS/MS data and can enable proteoform identification using discordant peptide abundance and proteogenomic search analyses. Typically, protein inference engines use peptide-level data to detect the presence or absence of peptides to identify protein isoforms. However, here we show the utility of incorporating quantitative profiles of peptides mapping to known isoforms in potentially increasing the sensitivity of proteoform detection. We thus hypothesized that previously generated LC-MS/MS plasma proteomic data can be reanalyzed at the peptide-level and using quantitative profiles to infer protein isoforms [20] and potentially yield deeper insights into putative disease mechanisms. Here, we then demonstrated a peptide-centric reanalysis of NP-based methodology upstream LC-MS/MS-based indicating known and novel disease-relevant proteoforms.

We performed peptide analysis using DIA data derived from healthy and early NSCLC subjects by conducting a discordant peptide intensity search to identify protein isoforms. We identified four proteins with putative isoforms, including BMP1, C4A, C1R, and LDHB. Importantly, none of these proteins showed a difference in abundance at the protein-level. For BMP1 and C1R, using peptide abundance as a proxy for functionally relevant protein we identified potential NSCLC-related isoforms. BMP1 is known to act as both a suppressor and activator, a function we show can be linked to differential abundance of two isoforms (long and short). Additionally, C4A showed distinct peptide abundance discordance in one segment of the protein, which did not correspond to any known protein coding isoforms, suggesting peptide-centric proteoform identification may result in novel disease-associated isoforms.

The method we used to search for protein isoforms through discordant peptide intensity is rather stringent in terms of the number of protein isoform candidates we can find but easily interpretable. Similar approaches such as COPF [33] and PeCorA [34] use quantitative disagreements between peptides mapped to the same protein or peptide correlation within the same protein to detect protein isoforms and suggest proteoforms. However, as shown with our examples where 2 of the 4 isoform candidates (C1R and LDHB) met the discordant peptide intensity criteria but failed to be readily explained by known isoforms or biological conjecture. Evaluation of the validity of the isoform candidate is needed but outside the scope if this study. For our examples, that process was mapping the peptides back to the genomic sequence and isoform transcripts. Manual validation (e.g., isoform specific enrichment with isoform specific antibodies) can confirm the presence of isoform. It was achievable for the four candidates we had from our isoform detection process, however, other processes such COPF and PeCorA could yield dramatically more candidates. A robust and automated evaluation method of true isoform is needed.

In addition, we showed that using subject-specific genotype data derived from WES can reveal subject-specific protein variants. Across 29 subjects we identified 422 protein variants, for which we observed peptides harboring SAAVs not present in standard peptide sequence search databases. Among these protein variants, we detected a protein harboring a genetic variant, rs1229984, with significant association with lung cancer, as well as cases where we observed both the reference and alternative alleles. Notably, low frequency alleles or those that significantly alter the physicochemical appearance of the protein's surface are broadly inaccessible for targeted affinity-based proteome screening tools (e.g., aptamer and antibody-based methodologies), demonstrating the unique synergy of NPs-based proteomics workflows coupled to unbiased LC-MS/MS readouts.

It is possible that the finding of four protein isoforms in 141 subjects and 422 protein variants in 29 subjects may have been impacted by limited sample sizes reducing our power to identify proteoforms. Similarly, it is also possible that other undiscovered proteoforms are not functional in plasma and may only be identified in other biofluids or tissues. While our study shows the utility of using NP-based methodology upstream LC-MS/MS-based workflows to identify proteoforms, it is possible that expanding the sample size and diversity in sample type may yield further insights into disease-associated proteoforms.

The identification of proteoforms (protein isoforms and protein variants) highlights important considerations for current approaches characterizing the impact of genetic variation on molecular phenotypes, like protein abundance, by conducting protein quantitative trait analyses (pQTLs). Specifically, recent pQTL analyses using large cohorts [35] are performed at the protein-level and largely miss or misattribute peptide-level proteoform effects. Furthermore, these studies utilize aptamer and antibody-based methodologies that, as been recently shown [35], can lead to false discoveries and uncertain identification error rates because of conceptual limitations (e.g., the presence of a non-synonymous SNP inducing an amino acid change that disrupts the binding of the aptamer or antibody). Together, this indicates that a large-scale LC-MS/MS-based peptide identification, proteoform identification, and pQTL study across multiple biofluids and tissues may be the next great frontier in proteomics.

Methods

NSCLC sample Collection

As part of an IRB-approved study, we collected plasma samples from subjects from 24 collection sites diagnosed with NSCLC at stage 1, 2, 3, and 4 post-diagnoses but before treatment, as well as samples from healthy and pulmonary comorbid subjects as controls. Diagnosis of NSCLC was based on CT-guided fine-needle aspirant biopsy. Respective collection site has IRB [20] approved protocols [20] and written informed consent from all subjects for obtaining blood samples from NSCLC subjects. For the healthy and pulmonary comorbid controls, subjects were enrolled based on call-backs at collection sites. Healthy controls did not have current diagnosis of any form of cancer or any pulmonary co-morbidities such as COPD or emphysema. All subjects were not necessarily fasted at the time of collection. For the plasma samples, they were collected in EDTA tubes, centrifuged, aspired, frozen, and stored at –70 °C within one hour of collection; Subsequent shipments of samples were on dry ice. Prior to Proteograph processing, plasma samples were thawed at 4 °C, aliquoted, and refrozen. Wilcoxon and Fisher tests on age and gender, respectively, did not show significant differences between control and NSCLC subjects.

DIA data generation

Peptides were reconstituted in a solution of 0.1% FA and 3% ACN spiked with 5 fmol/µL PepCalMix from SCIEX (Framingham, MA) for the SWATH-DIA analysis. We targeted a constant injection mass of 5 ug of peptides per 10 µL MS volume, but when lesser yield was observed, the maximum amount was injected. The mass spectrometer was operated in SWATH mode using 100 variable windows across the 400–1250 m/z range. We used a trap-and-elute configuration for each sample using an Eksigent nano-LC system coupled with a SCIEX Triple TOF 6600+ mass spectrometer equipped with OptiFlow source. Peptides were loaded on a trap column and separated on an Eksigent ChromXP analytical column (150 mm × 15 cm, C18, 3 mm, 120 Å) at a flow rate of 5 µL/min using a gradient of 3–32% solvent B (0.1% FA, 100% ACN) over 20 min, resulting in a 33 min total run time.

Peptide-Spectral Library generation

To build a peptide-spectral library, we used four plasma pools created from the patients in the lung cancer, depleted using a MARS-14 column (Agilent, Santa Clara, CA) and the Agilent 1260 Infinity II HPLC system, and analyzed by the Proteograph using the panel of 10 NPs. We used data-dependent mode on the UltiMate 3000 RSLCnano system coupled with Orbitrap Fusion Lumos using a gradient of 5–35% over 109 min, for a total run time of 125 min. To expand the spectral library, a separate pooled plasma consisting of 157 healthy and lung cancer subjects were also used, depleted using the MARS-14 column and fractionated into nine concatenated fractions with a high-pH fractionation method (XBridge BEH C18 column, Waters), and analyzed using the 10 NPs panel. Same DDA mode and parameters were used as the NSCLC samples. Finally, all DDA generated spectra were searched against human UniProt database using the Pulsar search engine in Spectronaut (Biognosys, Switzerland), and the final library was generated with a 1% FDR cutoff at the peptide and protein level.

DIA data processing

We used Spectronaut to process the SWATH-DIA data at default settings (version 13.8.190930.43655), with a Q-value cutoff at precursor and protein levels of 0.01 were used [20]. Per an ongoing, IRB-approved observational sample collection protocol, we have samples for 288 subjects as part of an NSCLC study assayed across a 10-nanoparticle panel. Subjects diagnosed with NSCLC stage 1, 2, and 3 were labeled as early NSCLC. Subjects with NSCLC stage 4 were labeled as Late NSCLC. In addition, we have healthy and pulmonary comorbid control arms. Subjects diagnosed with NSCLC but with Unknown stage were removed from analysis; subjects who did not have peptides detected in all nanoparticles in the 10-NP panels were also removed. Summary statistics of protein counts and peptide counts per protein were calculated at this point.

Next, proteins were filtered to those present in at least 50% of subjects from either heathy or early cases, leaving us with a total of 141 subjects (80 control and 61 early NSCLC). Peptide intensities were median normalized and natural logged.

Identification of protein isoforms

From the 1,565 proteins present after filtering, we searched for peptides that had differential abundance between controls and cancer ($p < 0.05$; Benjamini- Hochberg corrected). Discordant pairs are defined as peptides from the same protein where at least one peptide was identified with significantly higher, and another peptide was identified with significantly lower plasma abundance in healthy controls vs. early NSCLC.

Protein quantification across multiple samples

Within each nanoparticle, standard MaxLFQ was used to quantify abundance at the protein level. For each peptide, the intensity ratios between every pair of samples were first computed. The pairwise protein ratio is then defined as the median of the peptide ratios from all peptides map to the same protein. With all the pairwise protein ratios between any two samples, we can perform a least-squares analysis to reconstruct the abundance profile optimally satisfying all the protein ratios. Then the whole profile is rescaled to the cumulative intensity across samples for the final protein abundance [36]. A modified MaxLFQ was used to quantify abundance across samples and nanoparticles. For each protein, all peptides' intensities belonging to a protein from all samples and NP were employed to calculate peptide ratios and subsequent calculation steps resulting in abundance across all samples and NP.

WES data generation and processing

For 29 healthy, early NSCLC, late NSCLC, and comorbid subjects, DNA was extracted from buffy coats using QIAsymphony and eluted into 200 µl elution buffer. Sequencing libraries were then prepared using a TruSeq Exome Library Prep kit for each of the samples. Libraries were then sequenced on two lanes of NovaSeq S4 PE 101PE and received between 48M-88M reads and 9Gb-17Gb each. The bcl files were demultiplexed using BCL2FASTQ. We then used DRAGEN Host Software Version 07.021.510.3.5.7 and Bio-IT Processor Version 0x18101306 to call variants using the reference genome hg38 and generate a vcf file.

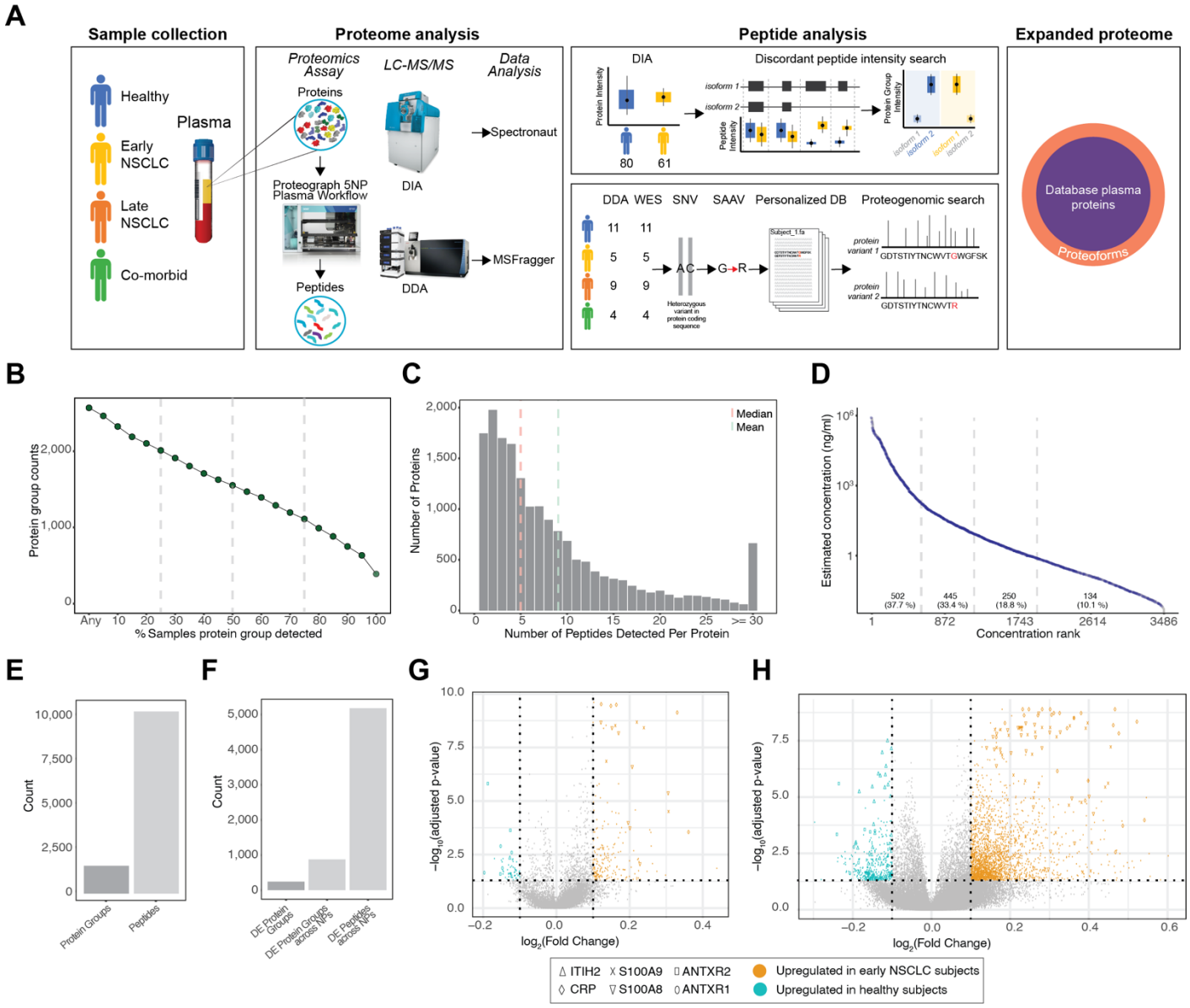
DDA data generation and protein variant identification

Similar to methods in [37], custom protein database was generated from human hg38 genomic FASTA, BED file from UniProt that describes that gene structure and VCF file from whole exome sequencing. Reference allele was generated using the FASTA file for nucleotide sequence and the BED file for the gene model with information on the location of the exons and the frame at which to translate the codons into amino acid sequence. For the alternative alleles, instead of generating an entire protein sequence, we generated tryptic peptides that span each specific mutation described in VCF file. If multiple variants are observed within a peptide, all possible combinations of the mutations are generated as peptides. MS/MS spectra from DDA data were searched against the custom protein database using the default Fragpipe pipeline (Fragpipe v15.0, Philosopher v.4.1.0 and MSFragger v3.4). For variant peptide identification, a 1% variant-peptide-level FDR was enforced using the target decoy approach [38].

Author information

MKRD and YH contributed equally. MKRD, YH, AS, JB, DH, SB, and OF conceived this study. MKRD and YH prepared the manuscript. MK, RB, TP processed the DIA and DDA data. JW and SK performed the proteogenomic analyses. YH, JB, and MKRD performed the protein isoform analyses. IM and DH supported LC-MS/MS data analyses. All authors discussed and reviewed the manuscript.

Figure 1: Proteome analysis of healthy and NSCLC subjects using a 5 NP plasma workflow



A. Overview of this proof-of-concept proteoform identification study. Plasma samples were collected from healthy (blue), early non-small cell lung cancer (NSCLC; yellow), late NSCLC (orange), and co-morbid (green) subjects (*Sample Collection*). The plasma proteomes were analyzed for each of these subjects, which included protein extraction, protein discovery using the NP-based Proteograph platform, then DIA AND DDA protein/peptide identification and quantification using LC-MS/MS and search algorithms (*Proteome Analysis*). Proteoforms were then identified using two strategies: 1) Discordant peptide intensity search, which included examining peptide mappings to known protein coding isoforms and using differential abundance to discover protein isoforms; and 2) Proteogenomic search, which included using genotype information (whole exome sequencing; WES) to perform personalized database searches and identify protein variants not captured in standard protein databases (*Proteoform Identification*). Together, these identified proteoforms represent an expanded plasma proteome database not captured in standard MS-based or targeted proteomic studies (*Expanded proteome*).

B. Dot plot representing the number of protein groups (y-axis) identified across study samples (x-axis), ranging from protein groups identified in one or more samples ("any") to proteins identified in 100% of samples ("100"). 25%, 50%, and 75% of samples are highlighted with grey dashed lines.

C. Bar plot showing the number of peptides detected per protein (x-axis) and the number of protein groups we observe for each bin (y-axis). The median (red dashed line) and mean (green dashed line) are shown.

D. Protein groups matching to a reference database (HPPP) are plotted as a distribution by the rank order of published concentrations (x-axis) and by the $\log_{10}$ published concentration (ng/ml; y-axis). The first, second, and third quantiles are highlighted with grey dashed lines. For each quantile, the number and fraction of protein groups matching the reference database are reported.

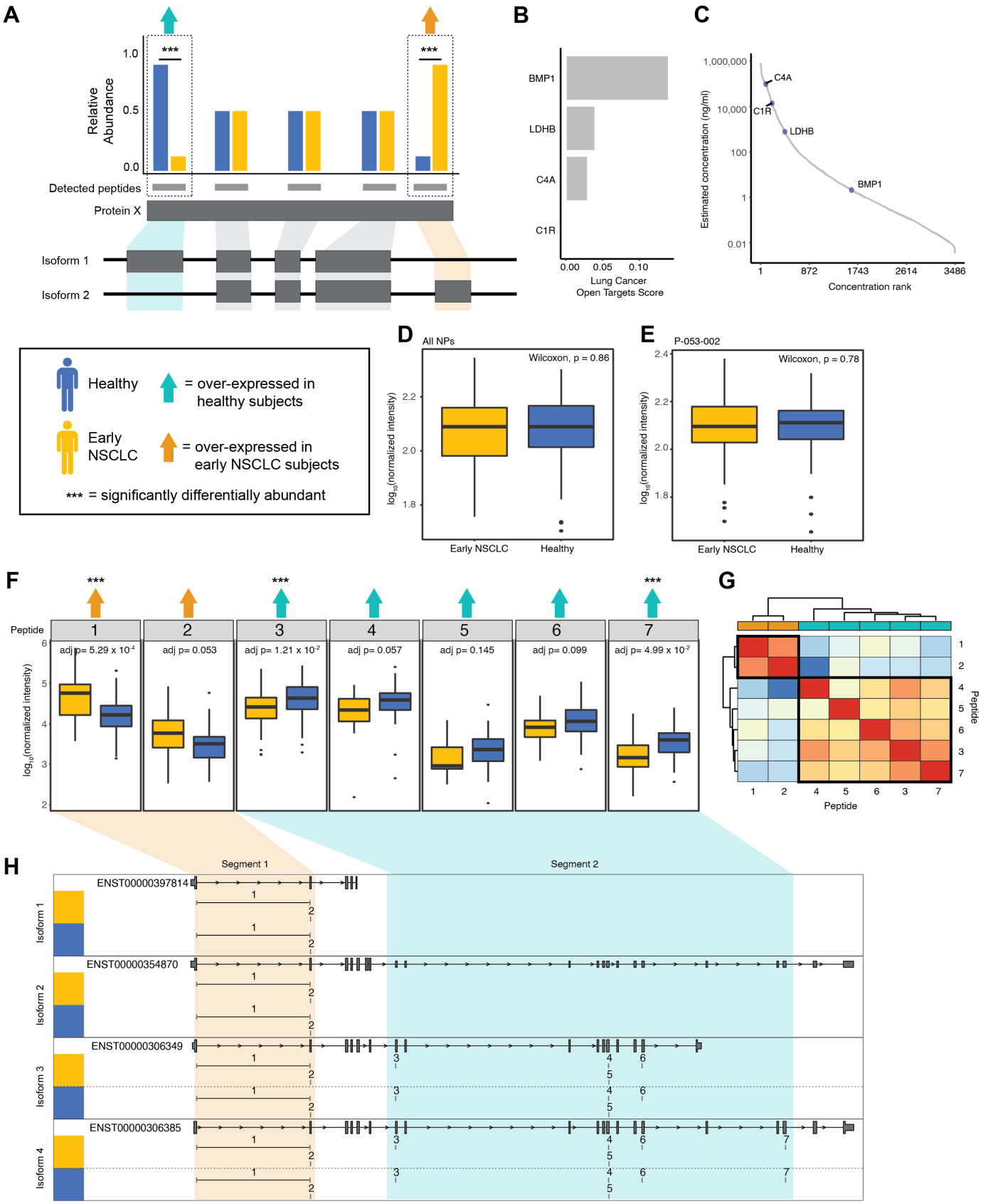
E. Bar plots showing the number of peptides and proteins groups retained after filtering to those present in at least 50% of subjects from either heathy or early NSCLC.

F. Bar plots showing the number of differentially abundant (DA): 1) protein groups, with collapsed abundances using MaxLFQ; 2) protein groups across NPs (i.e., DA independently across NPs); and 3) peptides across NPs.

G. Volcano plot showing the significance (adjusted p-value; y-axis) and fold change (x-axis) from calculating the differential abundance of protein groups across NPs between healthy and early NSCLC subjects. Protein groups with a log2(Fold Change) greater or less than 1.0 and adjusted p-value < 0.05 are highlighted, where protein groups with increased abundance in early NSCLC subjects are shown in orange and protein groups with increased abundance in healthy subjects are shown in teal. Proteins with known roles in cancer and immune response (ITIH2, CRP, S100A9, S100A8, ANTXR2, and ANTXR1) are highlighted with various shapes.

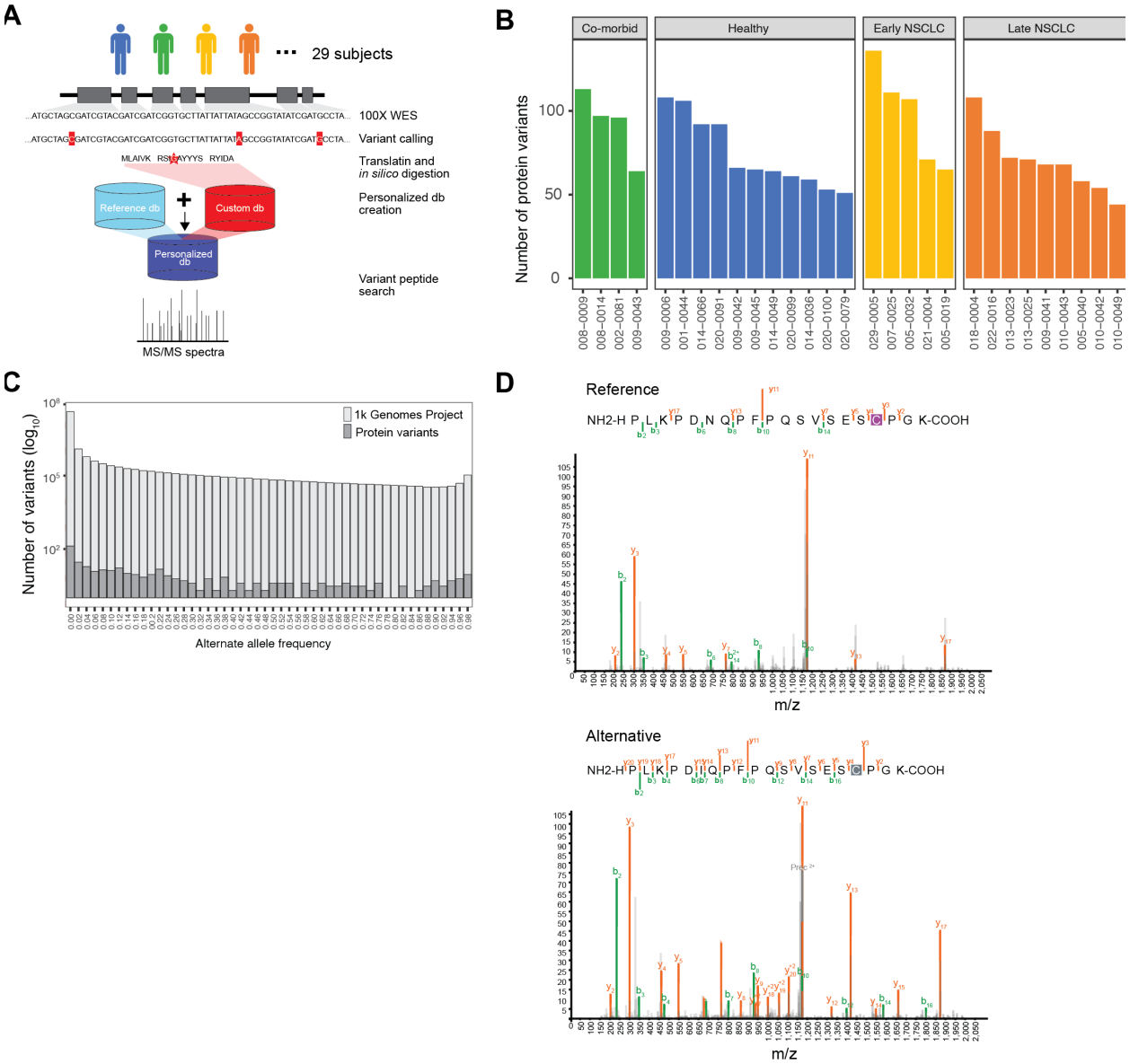
H. Volcano plot showing the significance (adjusted p-value; y-axis) and fold change (x-axis) from calculating the differential abundance of peptides across NPs between healthy and early NSCLC subjects. Peptides with a log2(Fold Change) greater or less than 1.0 and adjusted p-value < 0.05 are highlighted, where peptides with increased abundance in early NSCLC subjects are shown in orange and peptides with increased abundance in healthy subjects are shown in teal. Peptides mapping to proteins with known roles in cancer and immune response (ITIH2, CRP, S100A9, S100A8, ANTXR2, and ANTXR1) are highlighted with various shapes.

Figure 2: Identification of four proteoforms, including BMP1, in 141 healthy and early NSCLC subjects using a discordant peptide intensity search



A. Cartoon describing the discordant peptide intensity search strategy. We calculated DA across peptides between healthy (blue) and early NSCLC (yellow). Protein groups with at least one peptide significantly over-expressed (triple asterisks) in healthy subjects (teal arrow) and at least one peptide over-expressed in early NSCLC subjects (orange arrow) were identified as having putative proteoforms. Mapping the peptides to the gene structure, we inferred potential exon usage and segments suggesting the detection of more than one protein isoform.
B. Bar plot showing four proteins in which we potentially captured multiple protein isoforms: BMP1, C4A, C1R, and LDHB and their associated Open Target Score for lung carcinoma.
C. Plot showing the four proteins with putative proteoforms matched to a reference database (HPPP) plotted as a distribution by the rank order of published concentrations (x-axis) and by the $\log_{10}$ published concentration (ng/ml; y-axis).
D. Box plot showing the $\log_{10}$ median normalized intensities of BMP1 in early NSCLC subjects (yellow) and in healthy subjects (blue) with collapsed abundances across NPs. P-values, calculated using a Wilcoxon test, are shown.
E. Box plot showing the $\log_{10}$ median normalized intensities of BMP1 in early NSCLC subjects (yellow) and in healthy subjects (blue) in NP, P-053-02. P-values, calculated using a Wilcoxon test, are shown.
F. Series of boxplots showing the $\log_{10}$ median normalized intensities of seven peptides mapping BMP1 in early NSCLC (yellow) and healthy subjects (blue). Peptides that are over-expressed in healthy subjects are indicated with a teal arrow and in early NSCLC are indicated with an orange arrow. Peptides that are significantly DA are indicated with a triple asterisk. P-values, calculated using a Wilcoxon test and adjusted, are shown.
G. Heatmap showing the Pearson correlation of the seven BMP1 peptide abundances, where low correlation is indicated in shades of blue and high correlation is indicated in shades of red. Correlation values were clustered using hierarchical clustering. Peptides are annotated by the direction of DA, including over-expressed in healthy subjects are highlighted in teal and early NSCLC are highlighted in orange.
H. Gene structure plots of four known BMP1 protein coding transcripts (i.e., isoforms) with the seven BMP1 peptides mapped to genomic region. Peptides spanning intronic regions are indicated with a horizontal line. Peptides 1 and 2, corresponding to being over-expressed early NSCLC, are boxed in orange, creating one segment. Peptides 37, corresponding to being over-expressed healthy, are boxed in teal, creating a second segment. Segment 2 appears to correspond to the shorter isoform 1, whereas segment 2 appears to correspond to the longer isoforms 2-4.

Figure 3: Identification of 422 protein variants using a custom proteogenomic search



A. Cartoon describing a proteogenomic search to identify protein variants. Exomes of 29 subjects were sequenced to 100X, followed by genomic variant calling. The genomic sequence, including identified variants, were translated to protein sequence and digested *in silico*. These custom sequences were then combined with a reference sequence database to generate a personalized database. Personalized databases were generated for each of the 29 subjects and used to analyze the MS DDA data and search for variant peptides.

B. Bar plots showing the number of protein variants (y-axis) identified across 29 subjects (x-axis), including subjects with co-morbidities (green), healthy (blue), early NSCLC (yellow), and late NSCLC (orange).

C. Distribution of the number of variants (y-axis) across alternative allele frequencies (x-axis) in the 1k Genome Project and in the 422 protein variants.

D. Tandem mass spectra of peptides HPLKPDNQPFPQSVSESCPGK and HPLKPDIQPFPQSVSESCPGK arising from a heterozygous variant, where the alternative allele causes a single amino acid variant (SAAV; N→I). Both the reference peptide (top) and alternative (variant) peptide (bottom) are observed in the MS data.

References

1. Smith, L. M., Kelleher, N. L. & Proteomics, T. C. for T. D. Proteoform: a single term describing protein complexity. *Nat Methods* **10**, 186–187 (2014).
2. Li, Y. I. *et al.* RNA splicing is a primary link between genetic variation and disease. *Science (80-.).* **352**, 600–604 (2016).
3. Bogaert, A., Fernandez, E. & Gevaert, K. N-Terminal Proteoforms in Human Disease. *Trends Biochem. Sci.* **45**, 308–320 (2020).
4. Tucholski, T. *et al.* Distinct hypertrophic cardiomyopathy genotypes result in convergent sarcomeric proteoform profiles revealed by top-down proteomics. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 24691–24700 (2020).
5. Lisitsa, A., Moshkovskii, S., Chernobrovkin, A., Ponomarenko, E. & Archakov, A. Profiling proteoforms: promising follow-up of proteomics for biomarker discovery. *Expert Rev. Proteomics* **11**, 121–129 (2014).
6. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Publ. Gr.* **12**, 745 (2011).
7. Niroula, A. & Vihinen, M. How good are pathogenicity predictors in detecting benign variants? *PLoS Comput. Biol.* **15**, (2019).
8. Xiang, J. *et al.* Reinterpretation of common pathogenic variants in ClinVar revealed a high proportion of downgrades. *Sci. Rep.* **10**, (2020).
9. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).
10. Altshuler, D. L. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
11. Aebersold, R. *et al.* How many human proteoforms are there? *Nat. Chem. Biol.* **14**, 206–214 (2018).
12. Satpathy, S. *et al.* A proteogenomic portrait of lung squamous cell carcinoma. *Cell* **184**, 4348-4371.e40 (2021).
13. Kisluk, J., Ciborowski, M., Niemira, M., Kretowski, A. & Niklinski, J. Proteomics biomarkers for non-small cell lung cancer. *J. Pharm. Biomed. Anal.* **101**, 40–49 (2014).
14. Nishimura, T., Végvári, Á., Nakamura, H., Kato, H. & Saji, H. Mutant Proteomics of Lung Adenocarcinomas Harboring Different EGFR Mutations. *Front. Oncol.* **10**, 1–20 (2020).
15. Nishimura, T. *et al.* Current status of clinical proteogenomics in lung cancer. *Expert Rev. Proteomics* **16**, 761–772 (2019).
16. Anderson, N. L. The clinical plasma proteome: a survey of clinical assays for proteins in plasma and serum. *Clin. Chem.* **56**, 177–185 (2010).
17. Geyer, P. E., Holdt, L. M., Teupser, D. & Mann, M. Revisiting biomarker discovery by plasma proteomics. *Mol. Syst. Biol.* **13**, 942 (2017).
18. Geyer, P. E. *et al.* Plasma Proteome Profiling to Assess Human Health and Disease. *Cell Syst.* **2**, 185–195 (2016).
19. Bateman, A. *et al.* UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).

20. Blume, J. E. *et al.* Rapid, deep and precise profiling of the plasma proteome with multi-nanoparticle protein corona. *Nat. Commun.* **11**, 1–14 (2020).

21. Bach, D. H., Park, H. J. & Lee, S. K. The Dual Role of Bone Morphogenetic Proteins in Cancer. *Mol. Ther. - Oncolytics* **8**, 1–13 (2018).

22. Schwenk, J. M. *et al.* The Human Plasma Proteome Draft of 2017: Building on the Human Plasma PeptideAtlas from Mass Spectrometry and Complementary Assays. *J Proteom Res* **16**, 4299–4310 (2017).

23. Zhu, Y. *et al.* DEqMS: A method for accurate variance estimation in differential protein expression analysis. *Mol. Cell. Proteomics* **19**, 1047–1057 (2020).

24. Hamm, A. *et al.* Frequent expression loss of Inter-alpha-trypsin inhibitor heavy chain (ITIH) genes in multiple human solid tumors: A systematic expression analysis. *BMC Cancer* **8**, 1–15 (2008).

25. Ye, L., Sun, P.-H., Malik, M. F. A., Mason, M. D. & Jiang, W. G. Capillary morphogenesis gene 2 inhibits growth of breast cancer cells and is inversely correlated with the disease progression and prognosis. *J. Cancer Res. Clin. Oncol.* **140**, 957–967 (2014).

26. Rouleau, C. *et al.* The systemic administration of lethal toxin achieves a growth delay of human melanoma and neuroblastoma xenografts: Assessment of receptor contribution. *Int. J. Oncol.* **32**, 739–748 (2008).

27. Gebhardt, C., Németh, J., Angel, P. & Hess, J. S100A8 and S100A9 in inflammation and cancer. *Biochem. Pharmacol.* **72**, 1622–1631 (2006).

28. Watson, J., Salisbury, C., Banks, J., Whiting, P. & Hamilton, W. Predictive value of inflammatory markers for cancer diagnosis in primary care: a prospective cohort study using electronic health records. *Br. J. Cancer* **120**, 1045–1051 (2019).

29. Koscielny, G. *et al.* Open Targets: A platform for therapeutic target identification and Validation. *Nucleic Acids Res.* **45**, D985–D994 (2017).

30. Li, N. *et al.* Association between C4, C4A, and C4B copy number variations and susceptibility to autoimmune diseases: A meta-analysis. *Sci. Rep.* **7**, 1–9 (2017).

31. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

32. Govind, P., Pavethynath, S., Sawabe, M., Arai, T. & Muramatsu, M. Association between rs1229984 in ADH1B and cancer prevalence in a Japanese population. *Mol. Clin. Oncol.* **12**, 503–510 (2020).

33. Bludau, I. *et al.* Systematic detection of functional proteoform groups from bottom-up proteomic datasets. *Nat. Commun.* **12**, (2021).

34. Dermit, M., Peters-Clarke, T. M., Shishkova, E. & Meyer, J. G. Peptide Correlation Analysis (PeCorA) Reveals Differential Proteoform Regulation. *J. Proteome Res.* **20**, 1972–1980 (2021).

35. Pietzner, M. *et al.* Synergistic insights into human health from aptamer- and antibody-based proteomic profiling. *Nat. Commun.* **12**, (2021).

36. Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* **13**, 2513–2526 (2014).

37. Ruggles, K. V *et al.* An analysis of the sensitivity of proteogenomic mapping of somatic mutations and novel splicing events in cancer. *Mol. Cell. Proteomics* **15**, 1060–1071 (2016).

38. Elias, J. E. & Gygi, S. P. Target-Decoy Search Strategy for Mass Spectrometry-Based Proteomics. *Methods Mol. Biol.* **604**, 55–71 (2010).