INCORPORATION OF GLYCOPROTEOME DETECTION INTO LARGE SCALE UNBIASED PROTEOMICS STUDIES UTILIZING NANOPARTICLES

Bruce Wilcox[1], Kavya Swaminathan[1], John Blume[1], Jared Deyarmin[1], Preston Williams[1], Chinmay Belthangady[1], Manway Liu[1], Mi Yang[1], and Philip Ma[1]
[1]**PrognomiQ**, San Mateo, California, USA

INTRODUCTION

Challenges in early cancer detection have led to poor survival rates, highlighting the need for early diagnostic test development. Biomarkers measured in liquid biopsies offer a less invasive and accessible strategy for early cancer detection. Analyte degradation and dilution in complex biological matrix limit high specificity and sensitivity measurements, making biomarker discovery from blood a formidable challenge. PrognomiQ has developed a comprehensive multi-omics platform integrating multiple analyte measurements, cutting-edge analytical instrumentation, and novel data-analysis approaches.

Deep unbiased proteomics analysis in our platform is facilitated by recent advances in sample preparation (i.e. Seer's Proteograph™ Product Suite) coupled with improved mass spectrometry instrument sensitivity and speed. Together they provide the ability to quantify thousands of proteins from human plasma without compromising throughput or reproducibility.

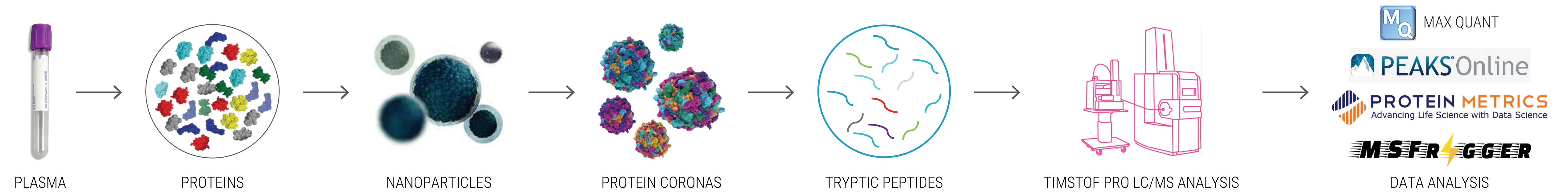
Additionally, we discovered that Seer's Proteograph™ nanoparticle technology specifically captures unique proteoforms in the corona formation and allow for comprehensive assessment of the circulating glycoproteome. ▪

EXPERIMENTS & METHODS

In a recent study to detect cancer related biomarkers, we analyzed 212 subject K2EDTA plasma samples (116 cancer subjects and 96 healthy control subjects), prospectively collected following an IRB approved protocol, on the Seer Proteograph™ platform using the standard five nanoparticle panel. Resulting peptides were analyzed on a Bruker timsTOF Pro mass spectrometer in data dependent (DDA-PASEF) mode coupled with a Dionex Utilimate 3000 LC generating a 60 min gradient on a 50cm uPAC pillar array column (Pharmafluidics).

Data was searched with MaxQuant[1] utilizing the following search parameters: 0.1% peptide/protein FDR search, default timsTOF parameters searched against complete UniProt SwissProt human proteome database with 50% reversed decoys and contaminants. All datafiles for each nanoparticle were searched together. Due to the large number of total datafiles (1,200) comprising all subjects and all nanoparticles, we only searched the datafiles for an individual nanoparticle across all 212 subjects due to computational limitations.

This dataset of 212 subjects was also searched for the presence of glycosylated proteins utilizing MSFragger2, PEAKS PTM implemented in PEAKS online (Bioinformatics Solutions) and Byonic3 (Protein Metrics) using default N-glycosylation parameters on each platform. All searches were performed against the Human Uniprot database with 50% reversed decoys and contaminants. All 1,200 datafiles were searched with MSFragger utilizing the published open search parameters2 for N-glycosylation and filtered at 1% FDR (peptide/protein). Complete data from nanoparticles 1 and 2 (480 datafiles) were also searched with PEAKS online and filtered at 1% FDR (PSM) and –log10p score>50. ▪



DETECTED PROTEIN GROUPS
Number of Protein Groups detected across all samples

DETECTED PEPTIDE COUNTS
Number of Peptides detected across all samples

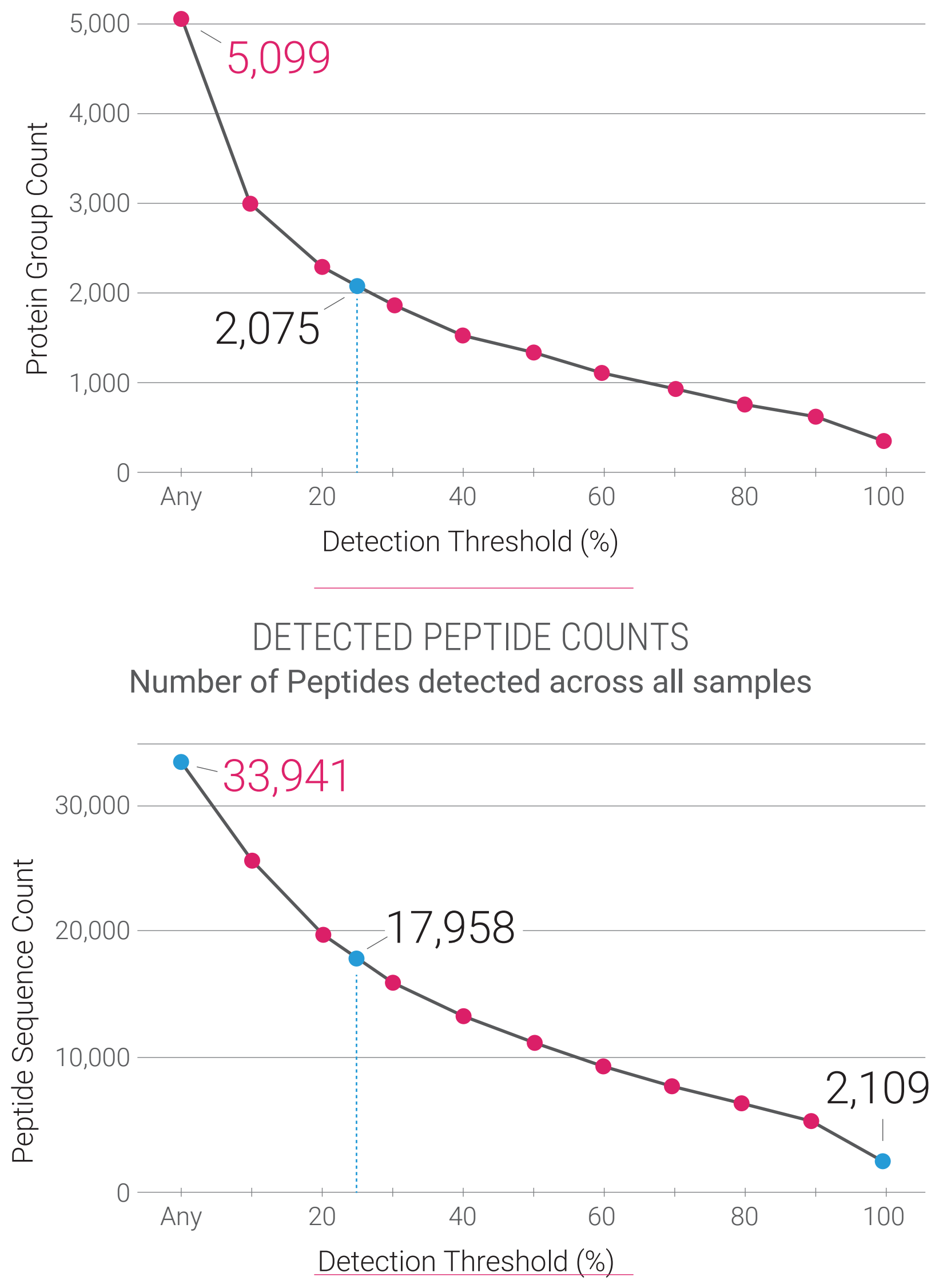


FIGURE 1

We detected 5,099 proteins groups and 33,941 peptides across all 5 nanoparticles for the 212 subject samples, with a median of 4 peptides per protein for proteins present in >25% of the samples. These results are comparable to previously reported results by Keshishian et. al4. utilizing depletion and fractionation but were generated in significantly less time per sample.

UNIQUE PROTEIN GROUPS
Number of Unique Protein Groups detected across all samples

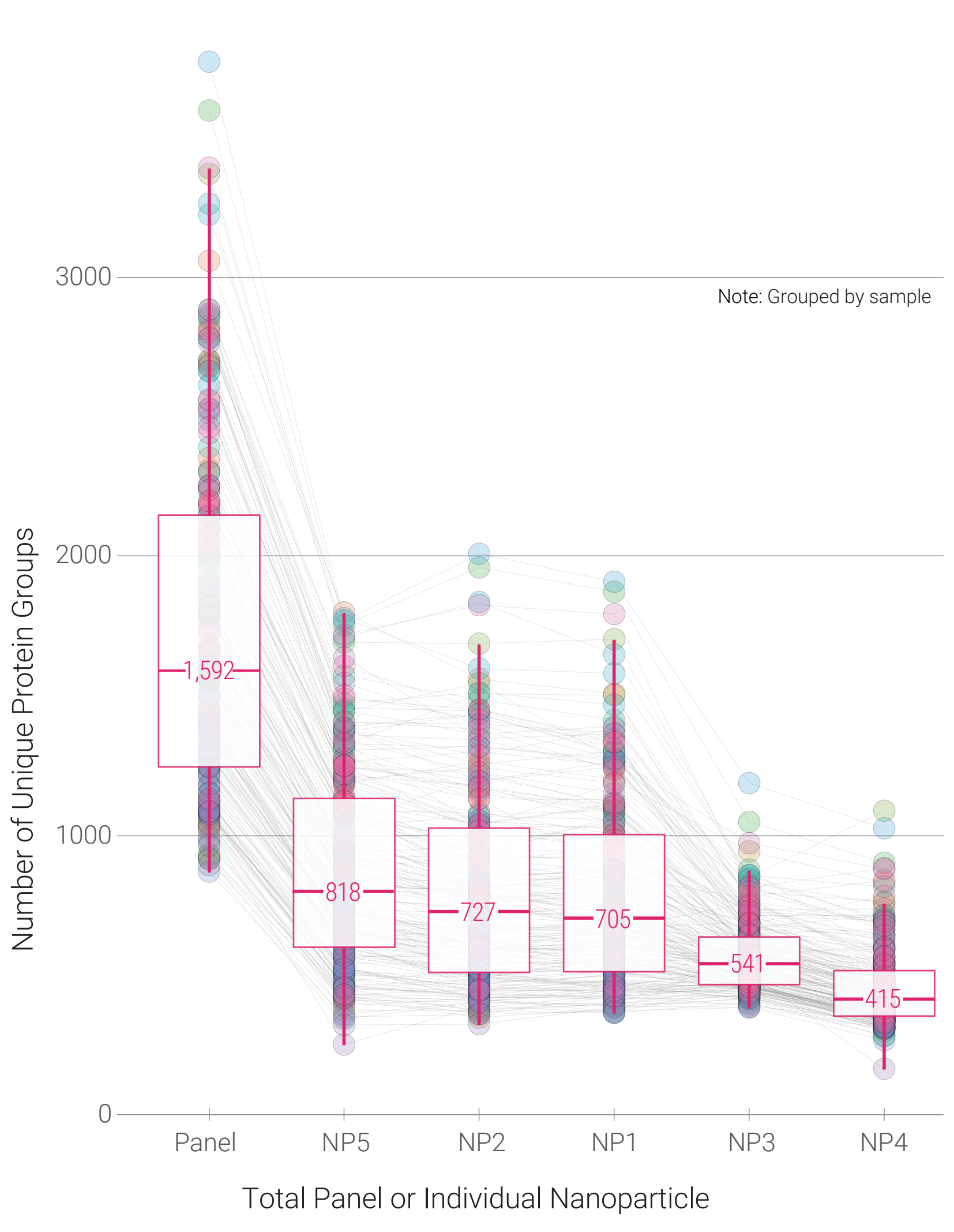


FIGURE 2

A median of 1,592 protein groups were detected across all 5 nanoparticles for all 212 subjects in this study, comparable to previously reported results by 5Blume, et al. NP5 provides the largest number, and most diverse, protein groups detected in any of the nanoparticles. Samples are grouped with connecting lines and colored by collections site. A high sample overall or for a given particle is generally then high in the other nanoparticles as well.

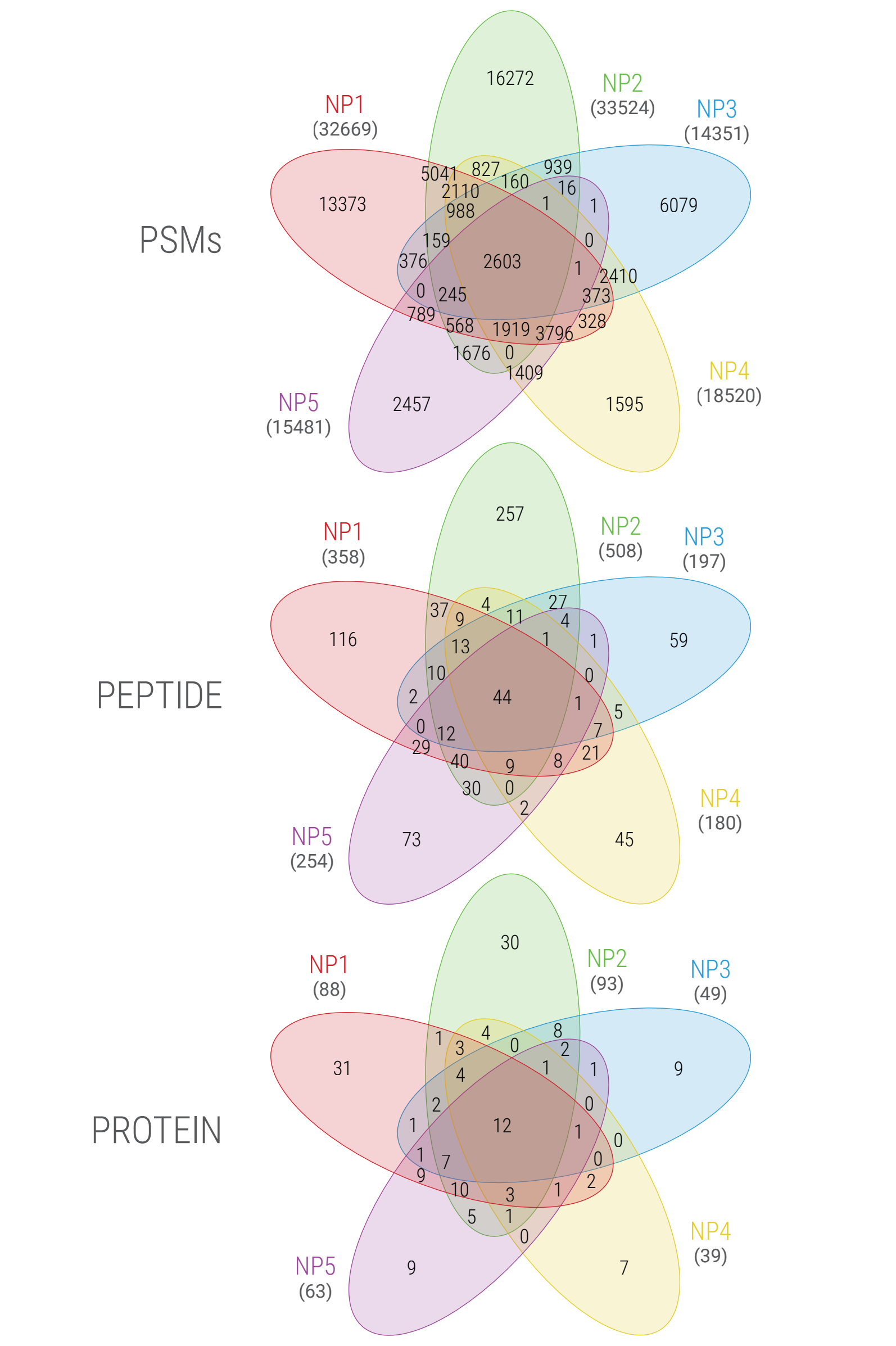


FIGURE 3

Individual nanoparticles yield both complementary and common glycoprotein identifications. (a) Peptide-spectral matches (PSMs) corresponding to (b) glycopeptides and (c) glycoproteins inferred from MSFragger searches across the five NP panel from ~1200 datafiles. A total of 66,511 glyco-PSMs from 877 unique peptides derived from 165 proteins were identified at 1% FDR, with NP1 and NP2 datasets accounting for ~80% of observed unique N-linked glycopeptides.

UNIQUE PROTEIN GROUPS
Number of Unique Protein Groups detected across all samples

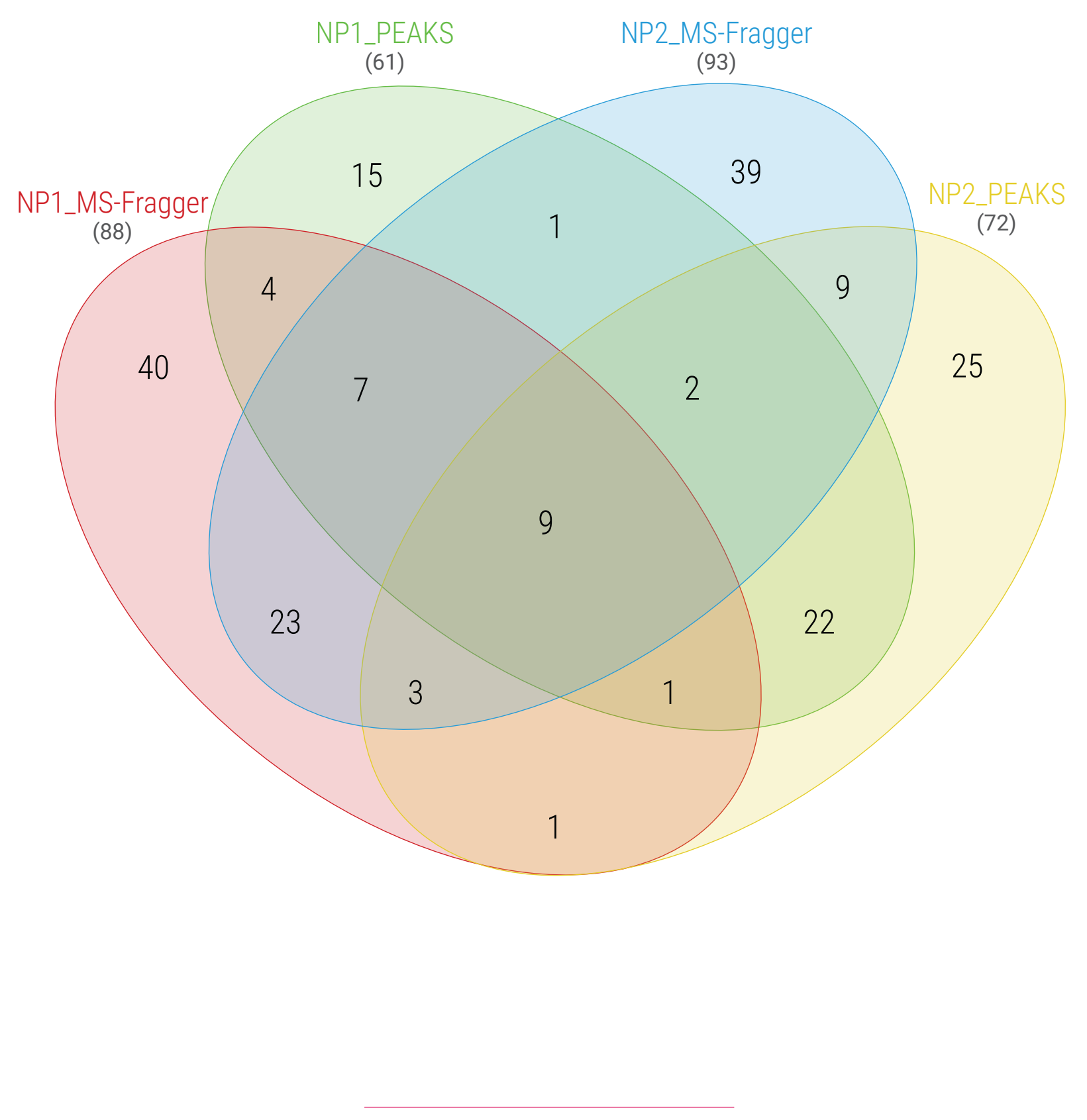


FIGURE 4

Venn diagram of proteins from which glycopeptides were detected in PEAKS or MSFragger glyco searches across nanoparticles NP1 and NP2 (480 files).

Approximately 75% (66/88) of proteins from which glycopeptides were identified via MSFragger and ~46% (28/61) of proteins with glycopeptides identified via PEAKS were also measured in the Max Quant label free quant (LFQ) search.

Interestingly, we observed little overlap between two algorithms at a glycopeptide level — three common peptides in NP1 and one in NP2 (data not shown). This is a likely result of high disparity in the repertoire of glycoforms searched by the two algorithms.

TOP 50 DETECTED PROTEINS GENECARDS CANCER SCORES

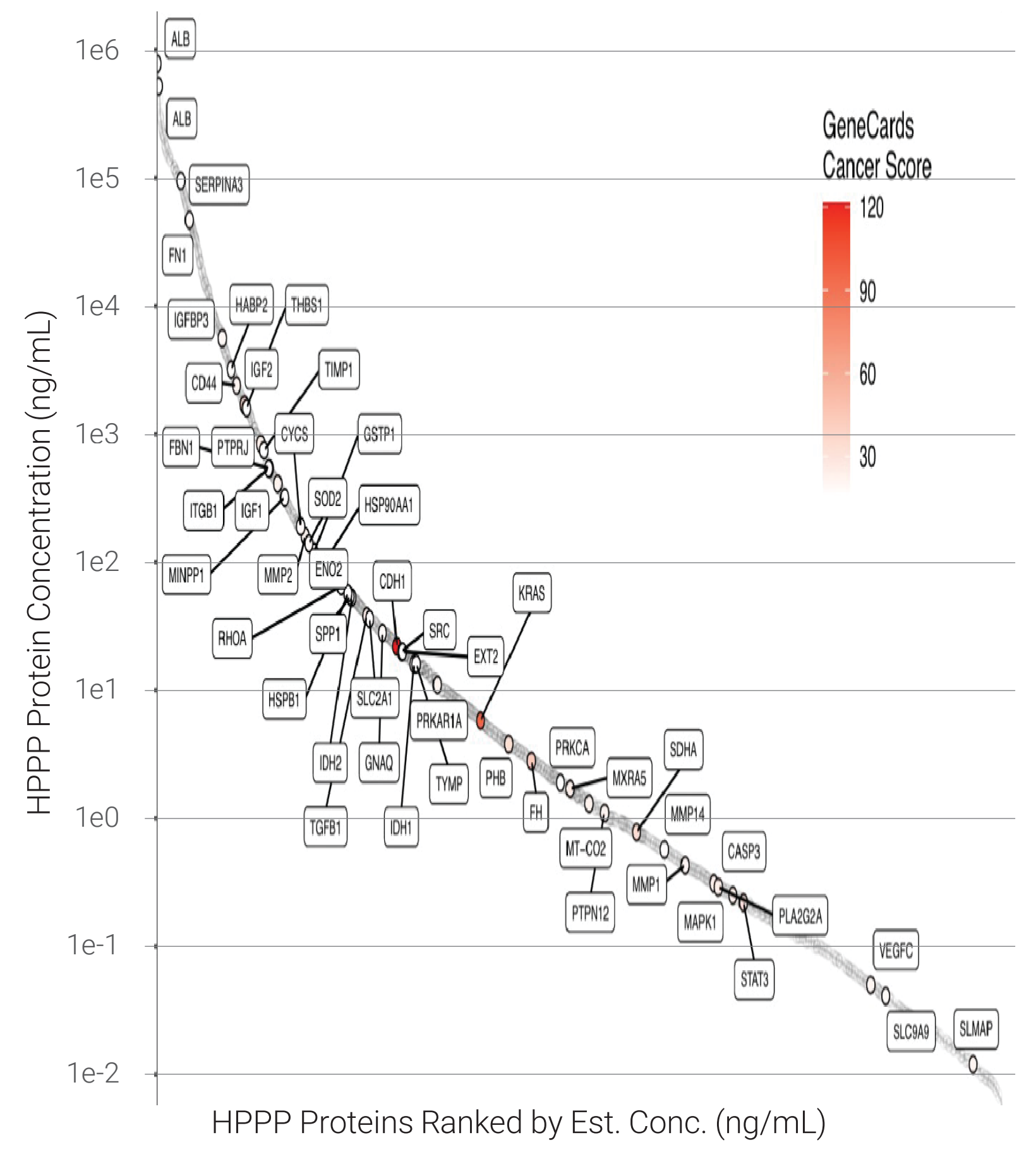


FIGURE 5

We mapped the detected 5,099 protein groups to the HPPP database, which illustrates the wide rage of reported protein concentrations (8 orders of magnitude) measured with the Proteograph nanoparticle technology and timsTOF Pro instrumentation. Additionally, we performed a Genecards6 analysis to determine the cancer associated proteins detected. The figure highlights the top 50 proteins, of which ~40% have a known plasma concentrations of <10ng/mL.

We found glycopeptides in PEAKS searches from 40 of the cancer associated proteins detected in the study. 42 cancer associated proteins detected in our study also had glycopeptides detected by MSFragger with agreement between the two algorithms on 16 overlapping glycoproteins. Of these 66 unique proteins for which glycopeptides were detected across both algorithms, 53 have multiple previously reported or predicted glycosites (Uniprot).

GLYCOSYLATED PEPTIDE MS/MS
Figure 6a

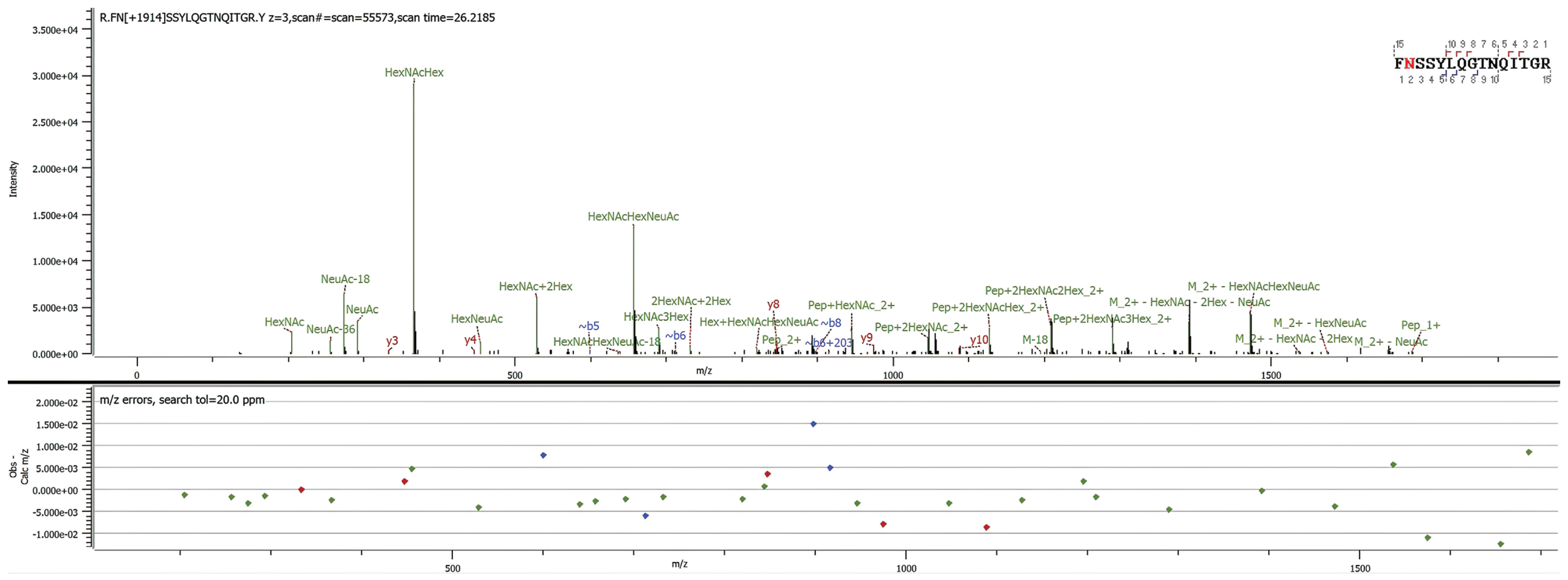


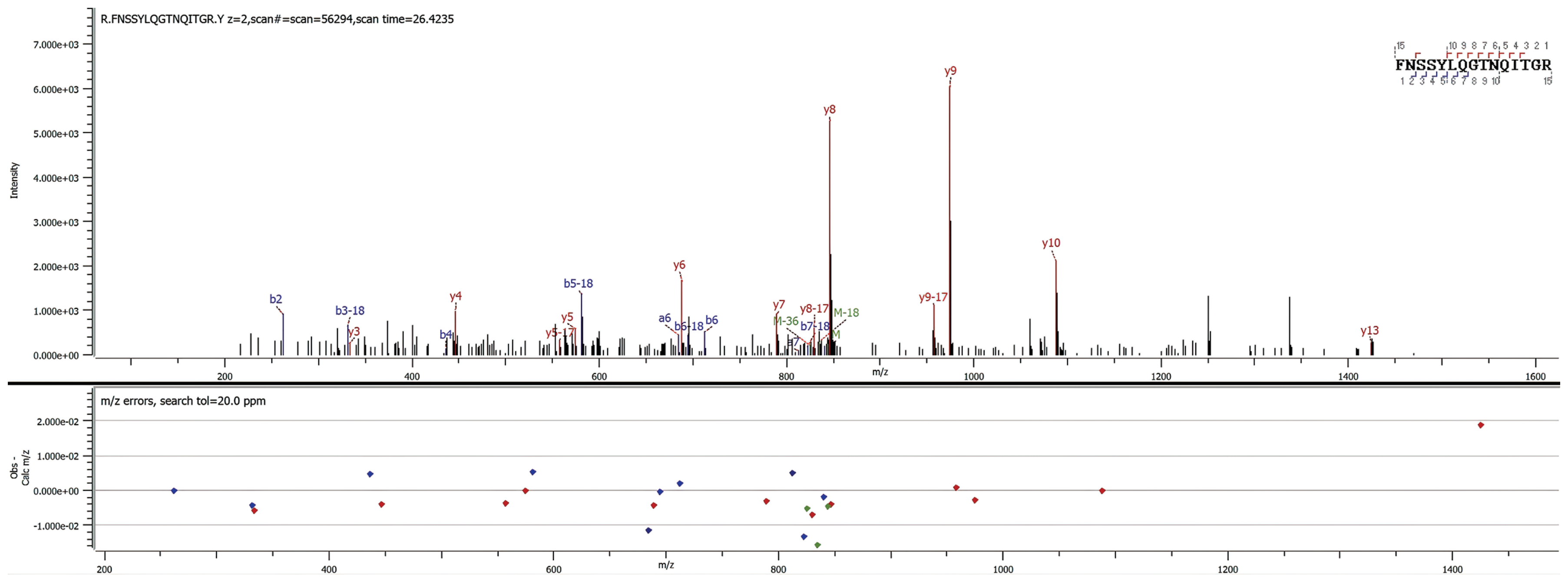
FIGURE 6

MS/MS spectrum from a (a) glycopeptide – FN(HexNAc-4Hex-5NeuAc-1)SSYLQGTNQITGR and (b) its unmodified counterpart, derived from Apolipoprotein-B (APOB), also shortlisted in the Genecard "cancer" search. The spectrum visualized from Byonic (Protein Metrics) following an N-glycan search of a single NP1 file.

The APOB glycosite captured by this peptide, N1523 has been previously reported7,8 in targeted glycoproteomics experiments deploying enrichment strategies.

Glyco-peptide ions along with N-acetylhexosamine (HexNAc) and N-acetylneuraminic acid (NeuAc) ions were detected in the MS/MS spectrum, consistent with the expected glycopeptide fragments expected to result from collision induced dissociation (CID).

UNMODIFIED PEPTIDE MS/MS
Figure 6b



CONCLUSIONS

- Unbiased proteomics of a large cohort of subjects utilizing Proteogragh and timsTOF technologies identified >5,000 proteins and 100's of glycosylated proteins in a single experiment. The simultaneous detection of native and glycosylated versions of the same proteins enables the analysis of differential expression of protein versions to detect associations with disease progression.
- Many of the glycosylated peptides detected are derived from known glycoproteins and have reported association with cancer.
- Glyco-searches with newly developed bioinformatics tools are computationally expensive, but do scale to >1,000 files. Each search engine detected complementary glyco-peptides/proteins that improved the overall detection rate and resulted in high confident identifications for the consensus identifications.
- Ongoing bioinformatics analysis and experiments to validate glycopeptides identified in this study.

REFERENCES

[1] Cox, J. and Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol, 2008, 26, pp 1367-72.

[2] Polasky, D.A., Yu, F., Teo, G.C., Nesvizhskii, A.I., Fast and Comprehensive N- and O-glycoproteomics analysis with MSFragger-Glyco. Nat Methods. 2020 November ; 17(11): 1125–1132

[3] Roushan, A. , Wilson, G.M. , Kletter, D., et al., Mol Cell Proteomics (2021) 20 100011

[4] Keshishian, H., Burgess, M., Specht, H. et al. Quantitative, multiplexed workflow for deep analysis of human blood plasma and biomarker discovery by mass spectrometry. Nat Protoc 12, 1683–1701 (2017).

[5] Blume, J.E., Manning, W.C., Troiano, G. et al. Rapid, deep and precise profiling of the plasma proteome with multi-nanoparticle protein corona. Nat Commun 11, 3662 (2020)

[6] Stelzer G, et. al The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analysis , Current Protocols in Bioinformatics(2016), 54:1.30.1 - 1.30.33.doi: 10.1002 / cpbi.5

[7] Liu, T. et. al Human plasma N-glycoproteome analysis by immunoaffinity subtraction, hydrazide chemistry, and mass spectrometry, J. Proteome Res. 4:2070-2080(2005)

[8] Chen, R. et. al Glycoproteomics analysis of human liver tissue by combination of multiple enzyme digestion and hydrazide chemistry, J. Proteome Res. 8:651-661(2009)

prognomiQ
www.prognomiq.com