Nanomics
The Next-Gen Proteomics Company

Use of the Proteonano™ Platform for Robust and Fast Large-cohort Proteomic Studies

Nanomics Biotechnology *haowu@nanomics.bio

Keywords:
Large-cohort Proteomics Studies, Proteonano™ Ultraplex Proteomics Platform, Data Standardization, Biomarker Discovery and Translation

Objective:
Based on the high-throughput ultraplex proteomics platform Proteonano™, we aim to achieve rapid and robust workflow for large-cohort proteomics studies.

I. Challenges in large-cohort proteomics studies

Generally, conducting large-cohort proteomics studies needs to process hundreds to thousands of clinical samples over dozens of experimental batches in a few months. This causes many pratical issues, including:

1.1 Sample preparation and handling issues:

- Sample quality: differences in sample origin and quality can affect the accuracy and reproducibility of protein detection.
- Consistency in sample processing: it's essential to ensure that all samples are processed in the same protocol (e.g., extraction, purification) to avoid systematic biases.

1.2 Methodology selection and validation:

- Choose the proper detection platform: select the mass spec or non-mass spec proteomics platform is crucial.
- Assay development: ensure that the chosen technology is effective for the sample types and experimental design and provides accurate, reliable data.

1.3 Data analysis and interpretation:

- Data preprocessing: efficiently handle and manage large volumes of raw data, including preprocessing, standardization, and normalization.
- Bioinformatics analysis: translate proteomics matrix data into biological insights, such as protein biomarker identification and quantification, pathway analysis.

1.4 Biomarker validation and reproducibility:

- Validation: verify key results using independent methods to ensure their reliability and reproducibility.
- Technical replicates: include technical replicates in your experimental design to evaluate data consistency and credibility.

1.5 Biological interpretation and functional analysis:

- Biological significance: connect proteomics data to biological functions and disease mechanisms, and conduct detailed functional and pathway analyses.
- Candidate protein identification: identify and validate candidate proteins or biomarkers with biological relevance.

1.6 Biomarker discovery and translation:

Translating multiplexed protein biomarker panels into clinical use often need customized assay development. Maintaining consistency between the discovery and translation phases is essential.

Furthermore, to handle these challenges effectivley, a team of experienced researchers, reliable technical platforms, and robust experimental quality control measures are essential to ensure robust and fast workflows for large cohort plasma proteomics studies.

"Robust and fast workflows are indispensable for successfully implementing large-cohort clinical plasma proteomics studies",

Professor Matthias Mann,
Max-Planck Institute of Biochemistry,
Proteomics and Signal Transduction
From www.evosep.com

II. Proteonano™ Ultraplex Proteomics Platform

The Proteonano™ Ultraplex Proteomics Platform, developed by Nanomics Biotechnology, is a highly automated and standardized system. It consists of the Proteonano™ Kit, Nanomics G1 workstation, and AI-powered proteomics analysis software, specifically addressing the bottlenecks in detecting low-abundance proteins in mass spectrometry-based proteomics. Proteonano™ Kit series is composed of AI-designed polypeptides to selectively bind and enrich low abundant proteins in biofluid samples with picogram sensitivity and peptide-level specificity.

Figure 1: Proteonano™ Ultraplex Proteomics Platform

Key Features	PGs identified	Sensitivity	Detection depth	Throughput	EU CE certified
	4000	pg / mL	9 logs	100SPD	

Proteonano™ quality control system (QCS)

In large cohort proteomics experiments with mass spec, subgroups of protein measurements with quantitatively different behaviors (i.e. batch effect) oftentimes hinders the downstream bioinformatic analysis. To address this issue, Nanomics' s Proteonano™ Ultraplex proteomics platform features a built-in quality control system (QCS), enabling comprehensive monitoring at the individual sample level. Here's how it works:

Step 1: Incubation controls

Proteonano™ Plasma Enrich Kit reagents are incubated with pooled human plasma samples (QC1 and QC2) to selectively bind low-abundance proteins. During this step, we monitor the protein intensities across multiple experimental batches by measuring the median coefficient of variation (%CV), which is to ensure the robustness of protein enrichment process driven by nano-bio interactions.

Step 2: Detection controls

Next, we conduct quality control on the peptides (QC3) enriched and purified from healthy human plasma using the Proteonano™ Plasma Kit. Since mass spectrometers perform variably over different manufacturers, conditions, and laboratory configurations, we monitor LC-MS/MS stability by assessing the reproducibility of peptide intensity across consecutive injections, thereby minimizing systematic differences due to instrument variability.

Step 3: Data controls

In this step, peptides are sequenced, assembled into proteins, and quantified using artificial intelligence algorithms. We also implement missing value imputation and normalization operations with widely accepted algorithms to assess and correct systematic variations across batches, ensuring consistent and reliable results.

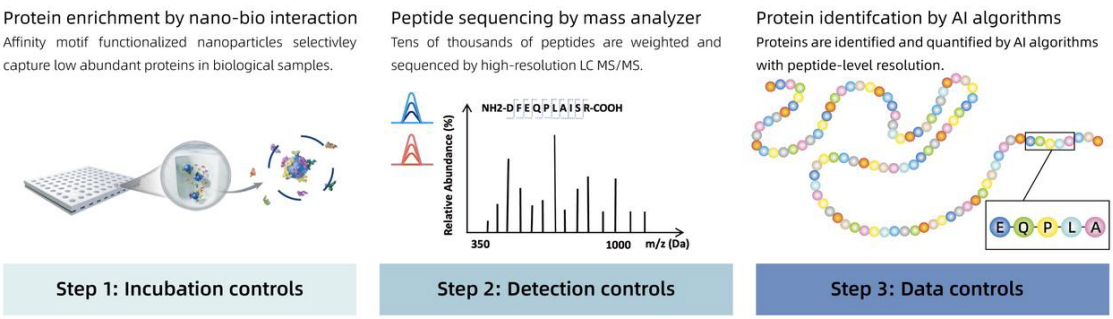


Figure 2: Proteonano™ platform built-in QC system

2.1 Incubation controls

For a standard 96-well plate experiment, we recommend using six quality control (QC) samples to monitor the entire platform's workflow, as shown in Figure 3. These include:

QC 1: QC1 is made up of pooled healthy human plasma samples to monitor the protein enrichment process (Step 1: Incubation controls) and correct potential inter-plate variations. Typically, three replicates of QC1 are included in each fully loaded 96-well plate.

QC 2: QC2 is also pooled healthy human plasma but remains untreated by the Proteonano™ Plasma Kit (i.e., neat plasma). It is used to monitor the steps of reduction, alkylation, enzymatic digestion, desalting, and lyophilization. Usually, one QC2 sample is included per 96-well plate.

QC 3: QC3 is a lyophilized peptide mix derived from pooled healthy human plasma that has undergone enrichment, reduction, alkylation, enzymatic digestion, and desalting. QC3 is used to monitor the performance of the LC-MS/MS instrument. When the LC-MS/MS meets baseline conditions, QC3 is injected three consecutive times to assess stability. QC3 sample is run after every 16-24 real samples to ensure consistent system performance.

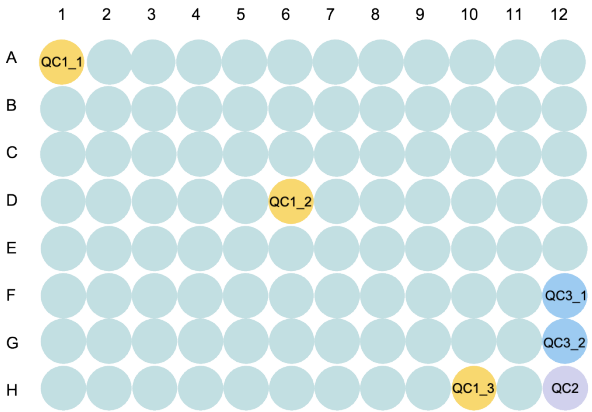


Figure 3: Recommended sample quality control solution

iRT

We add a standardized amount of iRT reagent (Biognosys) to each sample for calibrating and optimizing peptide quantification signals. This reagent includes 11 synthetic peptides that don't exist in nature, allowing us to monitor retention time, stability, and sensitivity under different chromatographic conditions. This aids in retention time calibration during database searches.

Since large cohort samples are often processed in multiple batches sampling, preprocessing, and analysis, adding a consistent amount of iRT to each sample provides a reliable reference for correcting batch effects across the entire project.

2.2 Detection controls

LC-MS/MS status check prior experiments

To evaluate the overall performance of the LC-MS/MS, we use QC 3 samples. These samples undergo the same procedures as the actual experimental samples, focusing on assessing the performance of the LC-MS/MS. Typically, a QC 3 sample is run every 10-20 LC-MS/MS runs. The quality control workflow for LC-MS/MS experiments is shown in Figure 4.

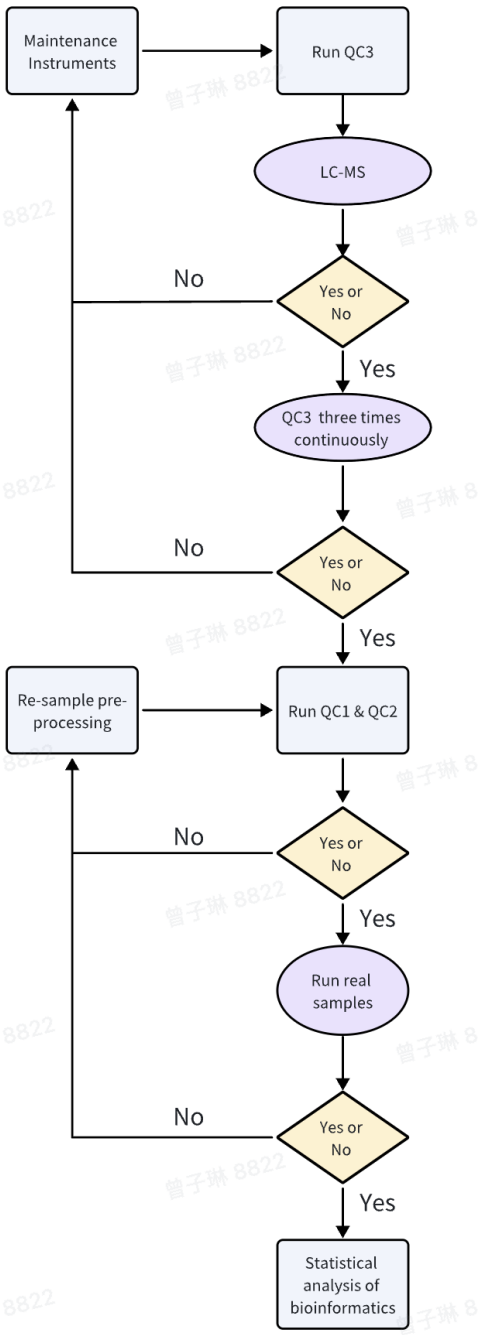


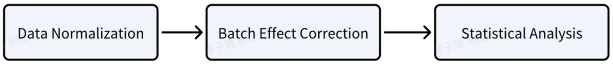
Figure 4: LC-MS/MS status check prior experiments

2.3 Data analysis controls

Batch effect correction

The development of high-throughput proteomics has made it possible to conduct experiments with tens of thousands of samples. But this scale inevitably introduces technical variability over subgroups, known as "batch effects". Batch effects can arise from differences in sample preparation, data acquisition conditions, and variations in technicians, reagent quality, and various instruments. These factors reduce the ability to detect genuine biological signals.

The Proteonano™ platform features an integrated data analysis workflow that ensures consistency of QC samples across multiple batches. It achieves a median interplate %CV of below 20% and an median intraplate %CV of below 15%.



TERMINOLOGY

The primary methods for diagnosing batch effects in proteomics data are [2] :

- **Principal Component Analysis （PCA)** identifies the main directions of variation (principal components) in the data. Projecting data onto these principal component axes allows for a visual assessment of sample similarity. By color-coding samples based on technical or biological factors, PCA helps identify what drives sample proximity. PCA is useful for examining clustering driven by biological or technical factors and for evaluating replicate similarity.

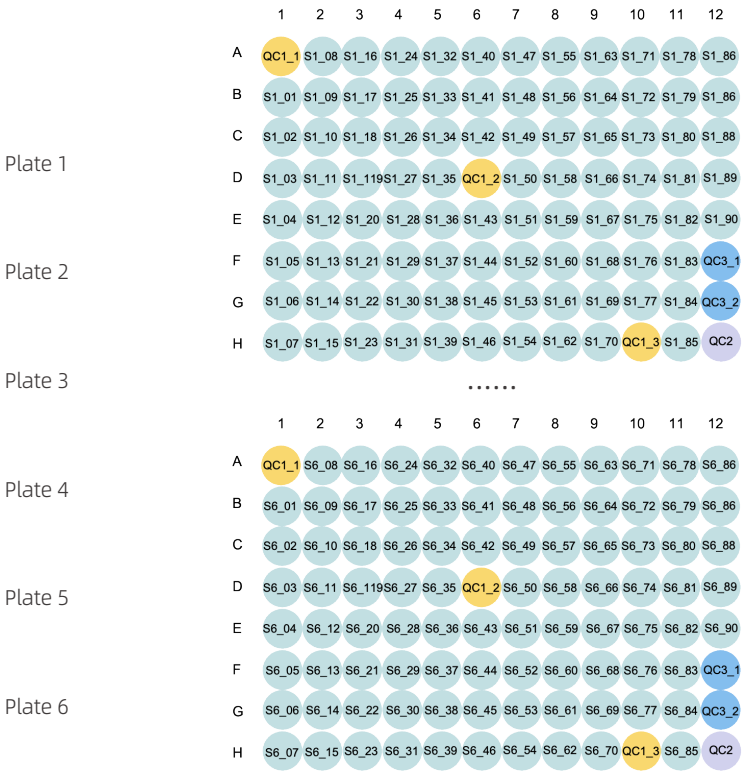
- **Hierarchical Clustering** is an algorithm that groups similar samples into a tree structure called a dendrogra m. By coloring the dendrogram according to technical and biological factors, you can easily see the reasons behind sample similarities. Hierarchical clustering is often used with heatmaps, which translate quantitative values into colors, making it easier to identify patterns in the data.

If similarity between samples is no longer influenced by technical factors, batch effects are considered corrected. In this case, neither PCA nor hierarchical clustering will show batch-specific clustering, and correlations among samples within the same batch will not be stronger than correlations among samples from different batches.

III. Case study: use the Proteonano™ platform to conduct a cohort with 540 plasma samples

Here’ s an example of how to quickly and robustly conduct a proteomics cohort study using the Proteonano™ platform with a cohort of 540 plasma samples. Following the experimental design from section 2.1, each 96-well plate is considered one batch. Each batch includes 3 QC1 samples, 1 QC2 sample, and 2 QC3 samples, with the remaining 90 wells used for the cohort samples. A total of 6 batches are required. The stability of the experiment is assessed by the reproducibility of the QC1 results. Intra-plate CV is calculated using the three QC1 samples in each batch, while inter-plate CV is derived from the 18 QC1 samples across all 6 batches.

3.1 Experimental design and sample configuration



Total samples: 540

Plate design: each 96-well plate is used for one batch

Total Plates: 6 plates

Per plate setup:

- 3 QC1 samples
- 1 QC2 sample
- 2 QC3 samples
- 90 cohort samples per plate

Figure 5: Experiment design and QC pLan

3.2 Stability of protein identification and inter-batch correlation for QC samples

We assessed the stability of protein enrichment using QC1 samples across 6 batches (a total of 18 QC1 samples). As shown in Figure 6, the median CV for protein quantity between batches is below 10%, and the median CV for protein intensity quantification is below 15%. This demonstrates the stability of protein enrichment across different batches in the project.

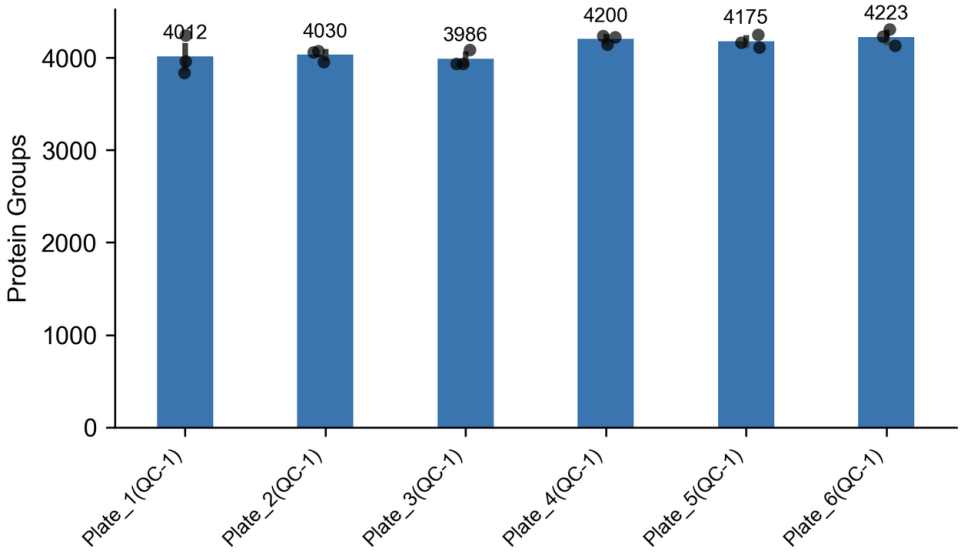


Figure 6: PGs identified for QC1 samples across 6 batches

We also assessed the pairwise correlations of the 18 QC1 samples throughout the project to ensure that Pearson correlation coefficients exceed 90% both between and within batches.

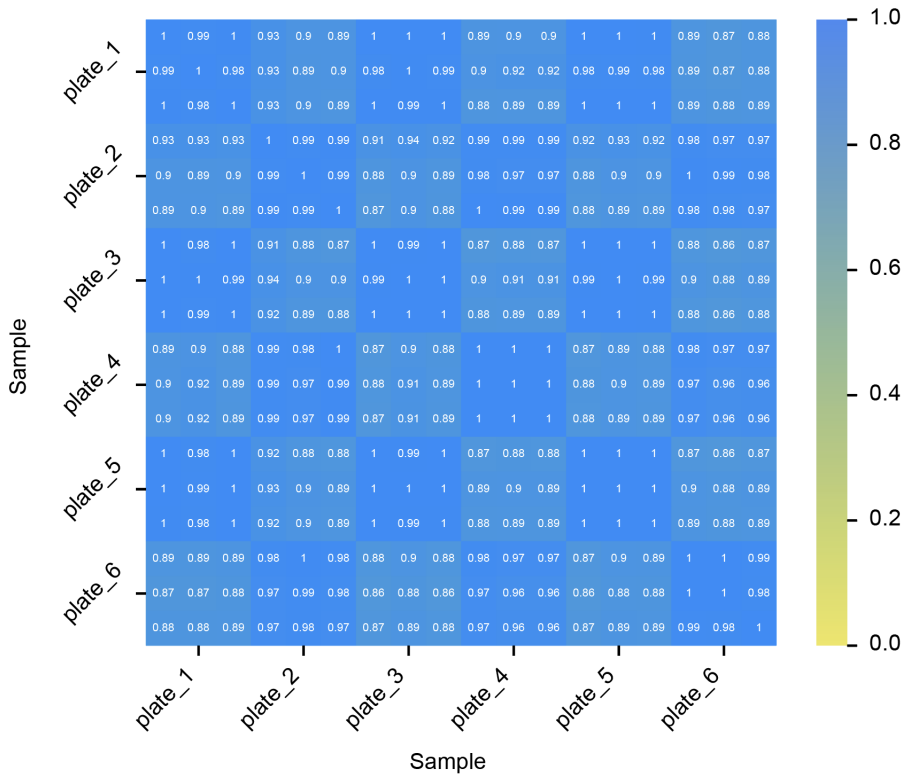


Figure 7: Pearson correlation across 18 QC1 samples from 6 batches

3.3 Batch normalization and batch effect correction

The intensity distribution of QC samples before normalization is shown in Figure 8.

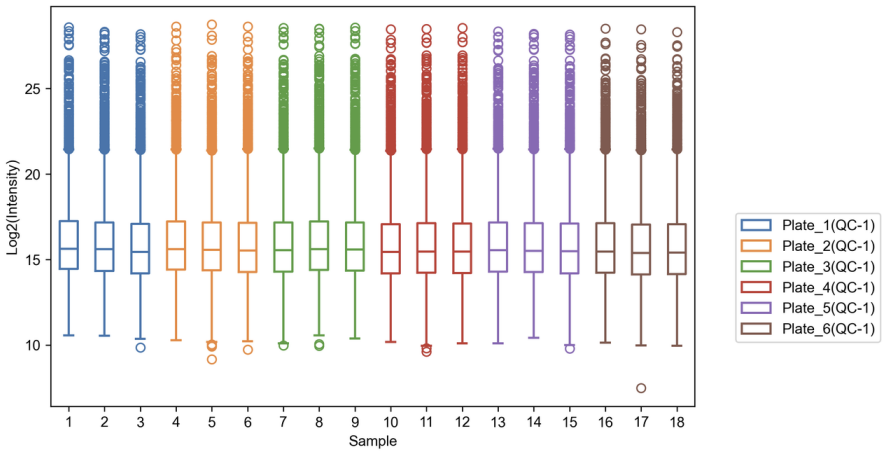


Figure 8: Protein intensity distribution before normalization

Variance-stabilizing normalization (VSN) was applied to the raw data, as shown in Figure 9. After processing, the median sample intensities become more consistent.

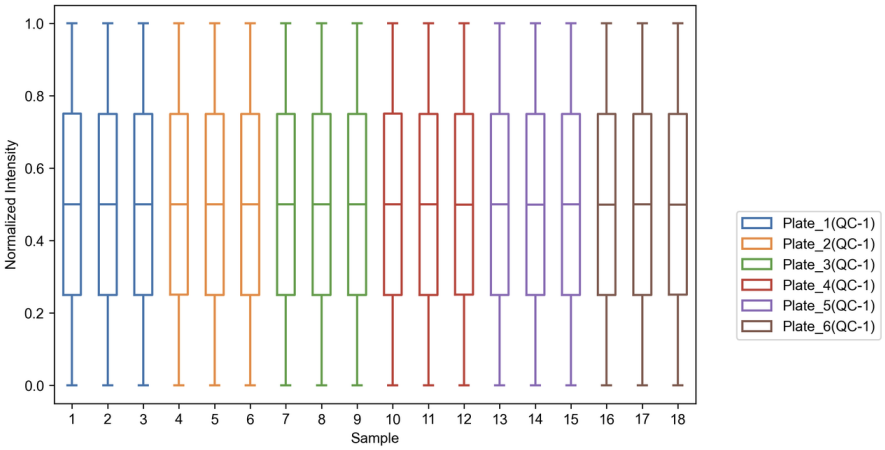


Figure 9: Protein intensity distribution after VSN

Batch correction was performed using ComBat[5], and the corrected median sample intensities are shown in Figure 10.

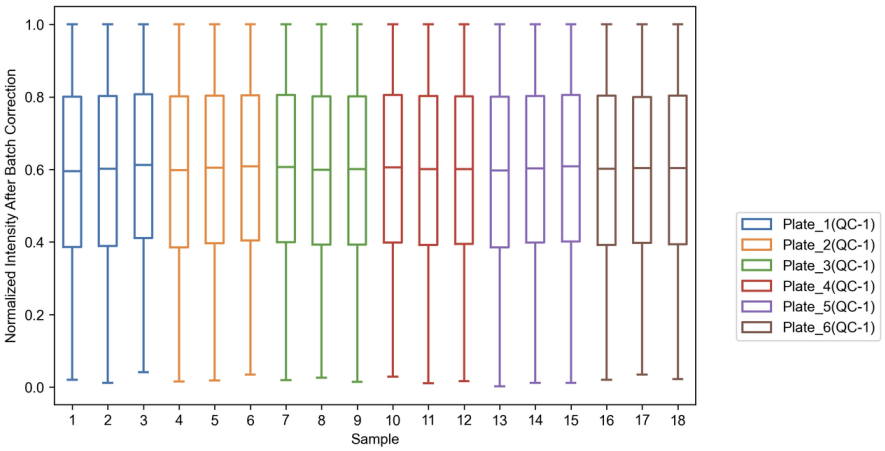


Figure 10: Protein intensity distribution after VSN and batch correction

3.4 PCA and clustering analysis

We conducted PCA (Figure 11) and clustering analysis (Figure 13) to evaluate sample similarity. The results revealed that QC1 samples from the same batch clustered together, while samples from different batches were distinctly separated.

After applying batch correction with ComBat [5], the PCA (Figure 12) and clustering analysis (Figure 14) showed significant improvement, with reduced separation between different batches and better clustering within the same batch.

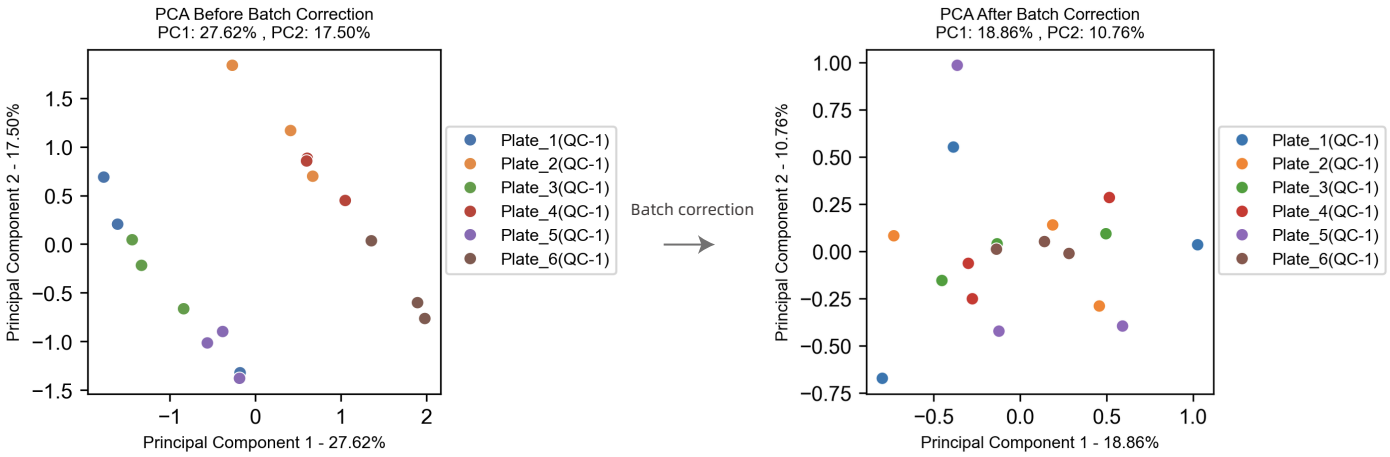


Figure 11: PCA distribution before batch correction

Figure 12: PCA distribution after batch correction

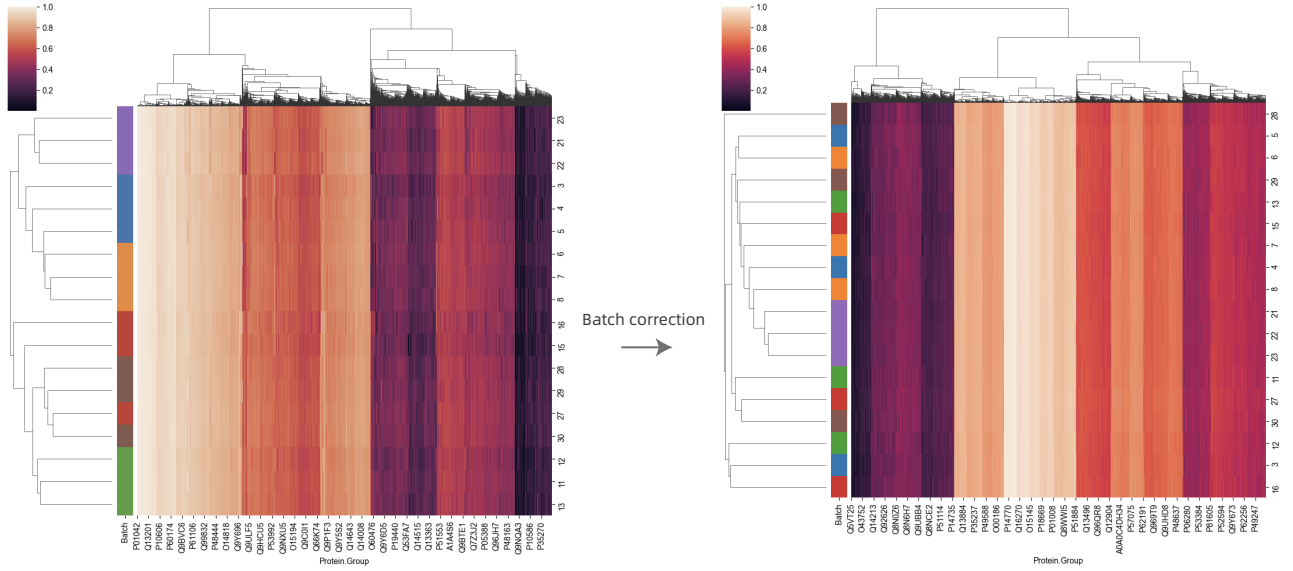


Figure 13: Sample clustering heatmap before batch correction

Figure 14: Sample clustering heatmap after batch correction

TERMINOLOGY

Coefficient of Variation（CV）

The coefficient of variation (CV) is a standardized metric used to assess variability in proteomics data. It is defined as the ratio of the standard deviation to the mean and usually expressed as a percentage. CV helps compare the variability of protein expression levels. In proteomics experiments, CV is commonly used to evaluate variability both between batches and within a batch.

Inter-plate CV

Inter-plate CV is a measure used to assess variability between different experimental batches. It evaluates consistency across replicate experiments performed in separate batches. The calculation of inter-plate CV is as follows:

$$\text{Inter-plate CV} = \left(\frac{\sigma_{\text{between plates}}}{\mu_{\text{overall}}} \right) \times 100\%$$

Here, $\sigma_{\text{between plates}}$ represents the standard deviation between different batches, while μ_{overall} is the 1overall mean across all batches.

Intra-plate CV

Intra-plate CV is used to assess variation within the same experimental batch. It helps evaluate consistency across repeated experiments within a single batch. The calculation method for intra-plate CV is as follows:

$$\text{Intra-plate CV} = \left(\frac{\sigma_{\text{within plate}}}{\mu_{\text{within plate}}} \right) \times 100\%$$

Here, $\sigma_{\text{within plate}}$ is the standard deviation within the same batch, while $\mu_{\text{within plate}}$ is the mean within the same batch.

3.5 Protein intensity CV values

The distributions of intra-plate and inter-plate CVs for protein intensity across the various QC samples are shown in Figures 15 and 16, respectively. The CV values are calculated as described earlier. After normalization, the median intra-plate CV for the six batches of QC1 samples is less than 10%.

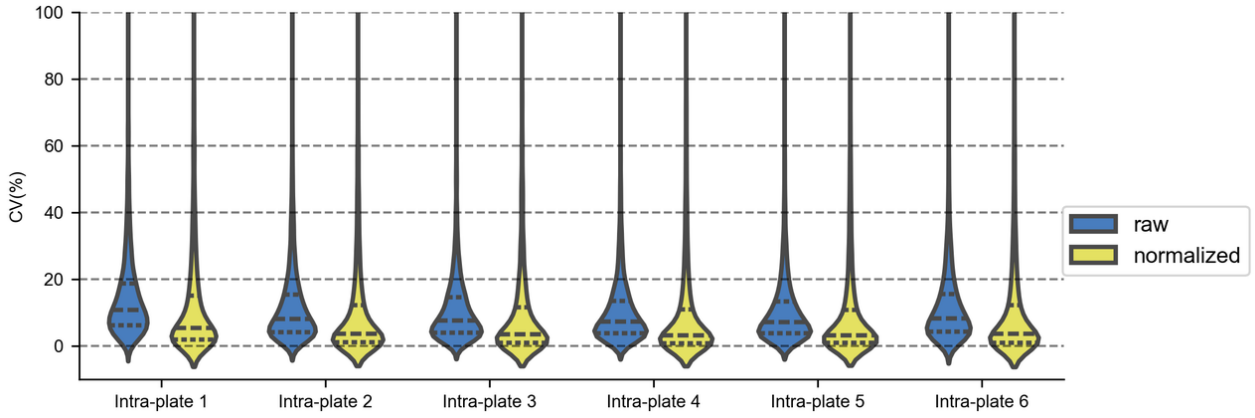


Figure 15: The Intra-plate CV for the six batches

After normalization, the median inter-plate CV for the six batches of QC1 samples is below 15%.

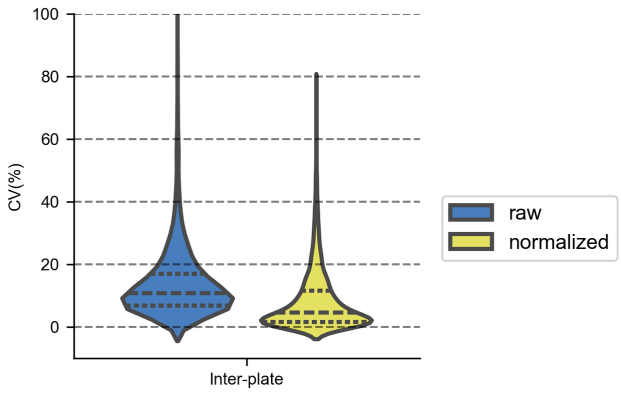


Figure 16: The Inter-plate CV for the six batches

Reference

1. Geyer PE, Voytik E, Treit PV, Doll S, Kleinhempel A, Niu L, Müller JB, Buchholtz ML, Bader JM, Teupser D, Holdt LM, Mann M. Plasma Proteome Profiling to detect and avoid sample-related biases in biomarker studies. EMBO Mol Med. 2019 Nov 7;11(11):e10427. doi: 10.15252/emmm.201910427. Epub 2019 Sep 30. PMID: 31566909; PMCID: PMC6835559.

2. Čuklina J, Lee CH, Williams EG, Sajic T, Collins BC, Rodríguez Martínez M, Sharma VS, Wendt F, Goetze S, Keele GR, Wollscheid B, Aebersold R, Pedrioli PGA. Diagnostics and correction of batch effects in large-scale proteomic studies: a tutorial. Mol Syst Biol. 2021 Aug;17(8):e10240. doi: 10.15252/msb.202110240. PMID: 34432947; PMCID: PMC8447595.

3. Wolfgang Huber, Anja von Heydebreck, Holger Sültmann, Annemarie Poustka, Martin Vingron, Variance stabilization applied to microarray data calibration and to the quantification of differential expression, Bioinformatics, Volume 18, Issue suppl_1, July 2002, Pages S96–S104, https://doi.org/10.1093/bioinformatics/18.suppl_1.S96

4. https://github.com/MannLabs/alphapeptstats/blob/main/alphastats/DataSet_Preprocess.py#L179

5. Behdenna A, Colange M, Haziza J, Gema A, Appé G, Azencott CA, Nordor A. pyComBat, a Python tool for batch effects correction in high-throughput molecular data using empirical Bayes methods. BMC Bioinformatics. 2023 Dec 7;24(1):459. doi: 10.1186/s12859-023-05578-5. PMID: 38057718; PMCID: PMC10701943.