

Best practices for addressing variability in experimental data collected with the Proteograph® Workflow

Introduction

The potential of liquid chromatography–mass spectrometry (LC–MS) proteomics to deliver biologic insight depends on well-designed studies that accurately detect and quantify a variety of proteins from vast datasets. As studies grow, researchers can clarify more subtle protein changes and biological signals to gain deeper insight. While deep proteomics studies are essential for refining our understanding of relevant molecular phenotypes, their inherent variability complicates the interpretation of LC–MS data, obscuring meaningful biological insights. The potential for technical variability grows as proteomics studies increase in size and complexity.

Here, we present Seer's core recommendations for addressing variability introduced during sample preparation, LC–MS acquisition, and multi-plate experiments. Our approach covers strategies for control-based performance assessment, normalization, and batch correction when handling data collected using the Proteograph workflow.

Performance assessment by assay control interpretation

Prior to MS data collection, biological samples must be processed to generate peptides for analysis, including extraction, alkylation, reduction, digestion, and cleanup/desalting. Sample processing itself can create reproducibility risks that can obscure findings or create artifacts,¹ especially when done manually.²

This technical note addresses:

- Bottom-up liquid chromatography–mass spectrometry (LC–MS) proteomics using data-independent acquisition (DIA)
- Seer's recommendations for data processing using the Proteograph Analysis Suite (PAS)
- Bioinformatic rationale for performance assessments, normalization, and batch correction

Although the Proteograph Product Suite includes measures to reduce variability through standardizing and automating sample processing, no workflow can eliminate it entirely. However, the strategic use of control samples at each key step in sample processing helps users untangle the various sources of noise in the data and potentially correct for them.

To support this, Proteograph assay kits include assay control samples for evaluating each sample preparation step of the automated assay workflows to identify the sources of potential process failures. Those controls are:

1. **User control:** An available well for users to add their own control sample that is processed identically to the sample wells
2. **Process control:** Lyophilized, undigested plasma proteins that are reconstituted and processed identically to the sample wells
3. **Digestion control:** Lyophilized, undigested plasma proteins that are reconstituted and processed from digestion onward
4. **Cleanup particles control:** Lyophilized, bulk-digested plasma peptides processed through only the peptide cleanup steps

Data from the User and Process Controls should be analyzed to determine if a general plate-level processing failure occurred. Meanwhile, data from the Digestion and Cleanup Particles Controls can be assessed in conjunction with log files and Seer's field support team to diagnose processing issues at the digestion or cleanup phases, respectively. To provide users with a basic framework, Seer has determined expected peptide mass yield ranges ([Table 1](#)) to give researchers an estimate of expected peptide yield values.

Table 1. Expected ranges of peptide yield for assay controls

Control	Expected yield range
Process control	0.56–1.43 µg
Digestion control	0.82–1.88 µg
Cleanup particles control	4.35–7.53 µg

It's worth noting that the User and Process Controls can also be used to monitor overall performance on the MS instruments as well. However, performance is heavily dependent on specific instrument hardware, settings, and data processing. For that reason, global absolute metrics cannot be determined. To monitor MS performance, Seer recommends that users run enough plates to create a performance baseline and use standard statistical methods (e.g., process control charts, Levey-Jennings charts) to establish allowable ranges unique to their study.

In addition, for studies that span multiple assay plates executed across many weeks and/or months, Seer recommends that users include additional control samples beyond the built-in Proteograph assay kit controls to account for the other sources of technical variability such as sample collection, storage conditions, freeze-thaw cycles, and temperature fluctuations. Ideally, each plate includes replicates of biologically distinct pooled controls representative of the study's complexity (e.g., male and female or obese and lean) to allow for per-protein, plate-level batch effect correction.

Precursor-level data normalization

Effective data normalization is essential for ensuring that proteomic information collected across samples, cohorts, days, instruments, and users can be compared for the extraction of impactful biological details. Variation within proteomic datasets can originate from several workflow steps, but analytical variation in sample measurements caused by hardware-associated changes can be a less visible source of inconsistency in results.

In bottom-up LC-MS proteomics, proteins are enzymatically digested into peptides, separated via liquid chromatography, and then ionized (often via electrospray ionization) before injection into the MS instrument. Typically, peptide ions are then analyzed using tandem MS/MS, a two-stage MS method in which intact peptides, known as precursors, are first separated by their mass-to-charge ratio (m/z) in MS1 and then fragmented before detection in MS2.

During the LC-MS/MS process, numerous component changes, from temperature fluctuations to hardware wear and tear, can impact the raw measured intensity of a peptide or precursor. This results in observed changes that do not necessarily reflect biological reality, thereby compromising data interpretation. For example, routine LC-MS column replacement can alter signal intensity, potentially leading to misinterpretation of biological signals between different samples.

Standard normalization approaches commonly subtract the median protein, peptide, or precursor intensity from the sample signal to address this type of signal variation.

- **Global median normalization:** A conventional approach, where the median intensity is estimated across all identified analytes in the sample
- **Density-filtered global median normalization:** A modification to the global median normalization that uses only precursors found in a certain fraction of the cohort (e.g., 80% of samples) to compute a global median that reduces outlier influence and improves accuracy in heterogeneous datasets
- **DIA-NN retention time dependent global normalization:** A variant of median normalization used in the MS data analysis package DIA-NN that computes several median normalization factors in retention time bins across the LC gradient to better account for elution-dependent effects on signal intensity
- **Peptide Calibrant-based ("PepCal") normalization:** A method developed at Seer that accounts for LC-MS drift over time using the median intensity of a set of calibration peptides with known retention times spiked into every sample at a consistent concentration.

Seer recommends:

Run enough plates to create a performance baseline and use standard statistical methods.

Include control samples beyond the built-in Proteograph assay kit controls to account for other sources of technical variability, ideally representative of the study's complexity.

Create a single large batch of calibrant-containing reconstitution buffer across an entire cohort study.

Use peptide calibrant-based normalization for most studies. Seer offers the Peptide Calibration kit and the PepCal normalization algorithm is available in PAS to facilitate proper normalization.

For most studies, Seer recommends using peptide calibrant-based (PepCal) normalization as it provides a known and reproducible way to estimate systematic LC-MS technical drift, independent of sample-specific effects that could adversely affect other normalization methods. PepCal normalization also addresses a common misconception that the majority of proteins (and thus the peptides) do not vary in concentration across different samples.

Numerous biological, pre-analytical, and analytical factors can affect protein concentration and prove this assumption false in plasma and other biofluid samples, especially as cohort studies increase in size and complexity. In fact, the Seer team has found that standard global median normalization can generate false differential protein abundance between samples, due to their composition. Since calibration peptides are added to samples at consistent concentrations, sample composition does not impact PepCal normalization.

To facilitate PepCal normalization, Seer offers the Peptide Calibration Kit that includes synthetic calibration peptides that can be added to samples during peptide reconstitution. To further improve PepCal normalization consistency, Seer recommends that users create a large batch of calibrant-containing reconstitution buffer that can be used across an entire cohort study. Additionally, the Proteograph Analysis Suite (PAS) includes a PepCal normalization protocol specifically designed for samples treated using the Peptide Calibration Kit.

If the Peptide Calibration Kit was not used in the preparation of the samples, Seer recommends using density-filtered global median normalization using proteins identified in at least 80% of the samples, which is less impacted by variation in sample composition than standard global median normalization.

Protein group quantification and batch correction

While MS measures peptides and their fragmentation patterns, protein-level metrics simplify downstream analysis and help drive biomarker discovery. To identify and quantify the proteins in a sample, individual peptides must be associated with a specific protein (or proteoforms) and have their measured intensities combined, or “rolled up.”

Specifically, Seer recommends using the *MaxLFQ* algorithm³ (as implemented by the *R* package *iq*⁴) to roll-up precursor intensity measurements into protein group intensities (note: for reasons not discussed here, in bottom-up MS proteomics protein identifications are reported as “protein groups”). *MaxLFQ* uses the relative peptide intensities in a sample to construct the most probable protein profile within the sample.

Seer recommends:

Use the MaxLFQ algorithm to roll up precursor intensity measurements into protein group intensities.

Properly randomize study samples both within and across assay plates to minimize batch effects.

*Perform batch correction using the `removeBatchEffect` function in the *R* package *limma*.*

*PAS includes built-in capabilities for running both MaxLFQ through *iq* and `removeBatchEffect` through *limma* for faithfully executing both protein group quantification and batch correction, respectively.*

Seer uses *MaxLFQ* because it is particularly advantageous for scalable label-free quantification and well suited for large studies of complex MS datasets, including those with missing peptide data. Additionally, the *MaxLFQ* algorithm can be easily extended for use on multiple measurements of a single sample (e.g., with the two wells of the Proteograph XT Assay), where precursors observed from the separate wells are treated as distinct species.

Once protein groups are identified and quantified, researchers must correct for any batch effects to ensure consistency in sample sets spanning multiple assay runs. Although the Proteograph platform can help users minimize pre-analytical variability, potential batch effects introduced prior to or during the Proteograph assay at the plate level can persist. To correct for these effects, Seer employs a linear model to fit the full dataset such that the plate effect’s influence on the data can be removed from the final intensity value. Seer recommends the use of the `removeBatchEffect` function in the *R* package *limma* to perform batch correction.⁵

It is important to note that the plate-level batch effect correction, like any batch correction, can attenuate biological signal. To ensure biological signal is not impacted, Seer encourages users to randomize study samples both within and across assay plates. For example, case samples should not be run exclusively on one plate and control samples on another.

Recommendation summary

As proteomics studies increase in size and complexity, Seer recommends the following for addressing variability introduced during sample preparation, LC-MS acquisition, and multi-plate experiments:

- To control for pre-analytical variability and workflow failures, the Proteograph assay kits include assay controls. However, Seer also recommends that users add additional pooled controls that are representative of the study and sample complexity.
- To account for technical LC-MS performance changes in large studies across time, Seer recommends the use of spike-in calibration peptides from Seer's Peptide Calibration Kit to serve as the basis for run-to-run PepCal normalization.
- For studies completed without spike-in calibration peptides, Seer recommends using density-filtered global median normalization.
- To roll up precursor-level data into protein group intensity values, Seer recommends the use of the *MaxLFQ* algorithm.
- To account for plate-specific batch effects, Seer recommends using the *removeBatchEffect* function from the R package *limma*.
- It is important to properly randomize samples on a plate to minimize batch effects.

These data handling algorithms and techniques are supported in Seer's Proteograph Analysis Suite, enabling consistent and reproducible results from import to reporting. Visit pas.seer.com to preview the platform.

References

1. Geyer, P.E., et al. [Plasma Proteome Profiling to detect and avoid sample-related biases in biomarker studies](#). *EMBO Mol Med* 11: e10427 (2019).
2. Soni, R. K. [Frontiers in Plasma Proteome Profiling Platforms: Innovations and Applications](#). *Clinical Proteomics* 21:43 (2024).
3. Cox, J., et al. [Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ](#). *Molecular & Cellular Proteomics*, 13 (9): 2513–2526 (2014).
4. Pham, T.V., et al. [iq: an R package to estimate relative protein abundances from ion quantification in DIA-MS-based proteomics](#). *Bioinformatics*, 36(8): 2611–2613 (2020).
5. Ritchie, M. E. et al. [limma powers differential expression analyses for RNA-sequencing and microarray studies](#). *Nucleic Acids Research*, 43(7):e47 (2015)

Find out more at
seer.bio/products/proteograph-product-suite

