illumina®

Achieve accurate and reliable protein detection with the Illumina Protein Prep solution

A guide to QC and data normalization steps using the DRAGEN™ Protein Quantification secondary analysis pipeline

Reproducible results

Attain consistent protein detection across samples and plates

Autolaunch analysis

Generate high-quality output with DRAGEN secondary analysis

Integrated insights

Analyze normalized protein count files easily with Illumina Connected Multiomics

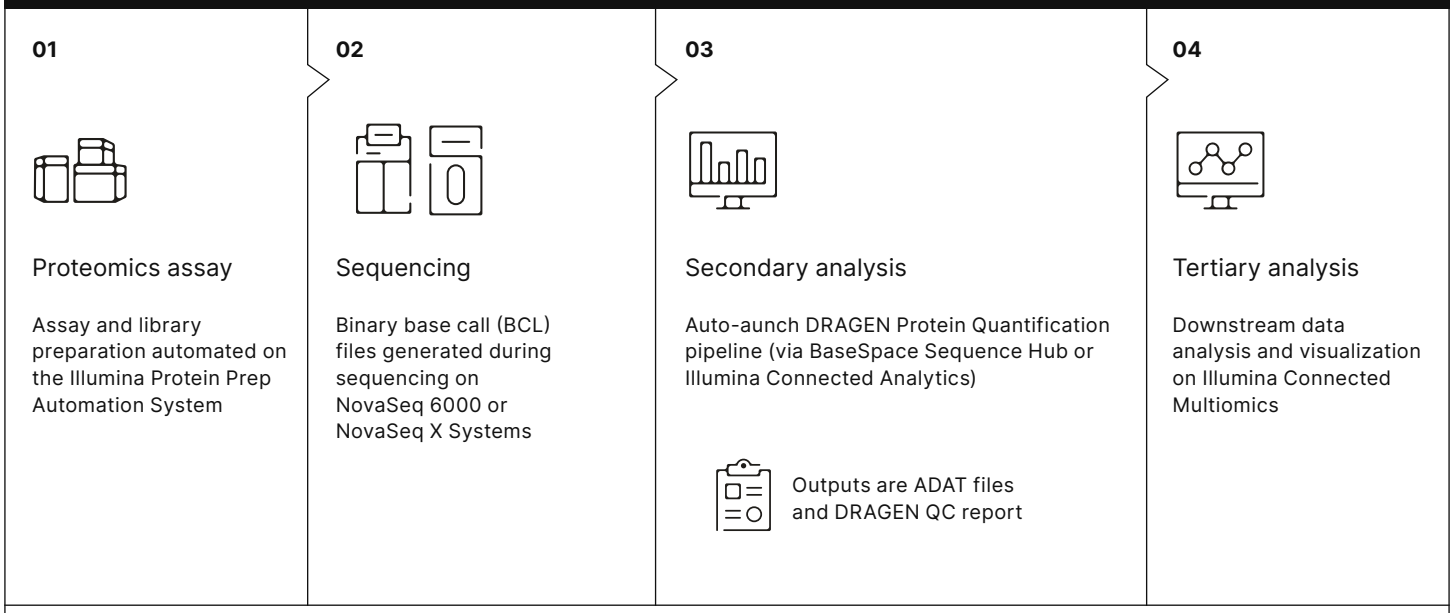
Introduction

Proteins play a key functional role in human biology, reflecting a real-time snapshot into health and disease states. Accurate and reproducible protein quantification is crucial for meaningful biological insights. Illumina Protein Prep is an innovative proteomics assay that uses SOMAmer® (slow off-rate modified aptamer) Reagents for protein capture, enabling highly sensitive protein detection. Combining this assay with next-generation sequencing (NGS) readout and the bioinformatics power of DRAGEN secondary analysis delivers a streamlined, high-throughput sample-to-results proteomics solution.

Illumina Protein Prep detects thousands of proteins simultaneously in a single sample, providing a powerful platform for biomarker discovery. However, plate-to-plate bias and technical variability of high-throughput proteomics platforms can obscure differential expression, particularly for biomarkers with small effect sizes. The DRAGEN Protein Quantification pipeline addresses this challenge during secondary analysis, with robust normalization methods to deliver high-quality, comparable data across experiments. On cloud, the pipeline launches automatically after sequencing is completed on a NovaSeq™ 6000 or NovaSeq X System and generates a DRAGEN QC report and ADAT files with normalized protein counts that can be used for downstream tertiary analysis with Illumina Connected Multiomics (Figure 1) to gain deeper biological insights into the proteome. The DRAGEN Protein Quantification pipeline can be installed on a DRAGEN on-premises server for users needing a local analysis option.

This technical note explains the quality control (QC) and normalization steps incorporated in the DRAGEN Protein Quantification pipeline to eliminate sample and technical variability for high-quality proteomics data.

Figure 1: An overview of the Illumina Protein Prep workflow.



The high-throughput proteomics solution uses an SOMAmer Reagent-based proteomics assay for sensitive protein detection from serum or plasma samples, followed by Illumina library preparation and sequencing on the NovaSeq 6000 or NovaSeq X Systems. The proteomics assay and library preparation workflow is automated on the Illumina Protein Prep Automation System, a custom Tecan Fluent 780, for consistent and reproducible results. DRAGEN Protein Quantification pipeline is auto-launched post-sequencing in BaseSpace Sequence Hub or Illumina Connected Analytics, transforming the proteomics data in BCL files produced by Illumina sequencing systems to normalized proteomics counts in an ADAT file with an accompanying DRAGEN QC report. The ADAT file can be integrated directly into tertiary analysis programs, including Illumina Connected Multiomics, for further data interpretation.

Why is data normalization needed?

Technical variability is inherent to any high-throughput proteomics assay and can arise from multiple sources, including sample preparation, intrinsic sample heterogeneity, hybridization efficiency, PCR efficiency, and sequencing. In addition to technical variability within a sample, batch effects can occur that affect all samples on the same plate, due to differences in operator, instrument, or environmental effects. Without effective normalization or batch correction, these systematic variations could obscure true biological differences, leading to inaccurate conclusions.

Data normalization is essential to correct for this variability, standardize data, and enable meaningful comparisons across samples, plates, and even different assay runs. Implementing comprehensive normalization strategies with the DRAGEN Protein Quantification pipeline provides reliable and reproducible protein quantification data.

Illumina Protein Prep controls

The Illumina Protein Prep assay includes plate controls and SOMAmer Reagent controls in each sample to mitigate assay variability. Both types of controls are used to generate normalized protein expression counts with the DRAGEN Protein Quantification pipeline.

Plate controls

Each 96-well plate accommodates 85 samples and 11 plate controls (Figure 2), including blanks, calibrator controls, and QC samples. These controls are included with the Illumina Protein Prep assay and are run in each plate along with the plasma or serum samples for normalization and QC. They serve as benchmarks to detect and correct for technical deviations and provide a foundation for aligning samples run on different plates.

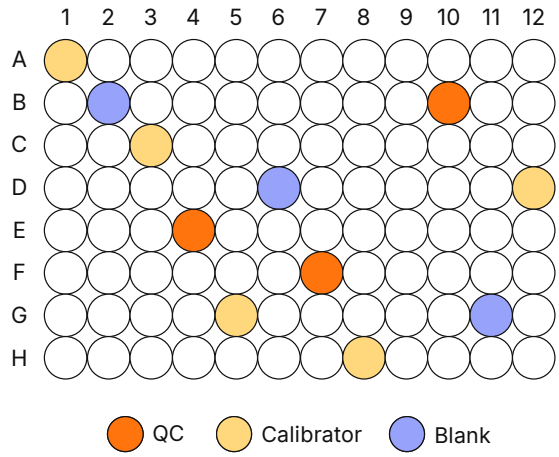
Blanks

Blank samples consist of pure buffer and provide an estimate of nonspecific background in the Illumina Protein Prep assay. These controls are expected to have significantly lower counts than plasma or serum samples. Elevated counts in blank wells could indicate that a run has elevated background or that sample contamination has occurred.

Calibrator controls

Calibrator controls allow for correction of plate-to-plate variability. These controls are either pooled human plasma or serum, depending on the sample types being run. Plate calibration normalization is performed by comparing the median of the calibrator controls per SOMAmer Reagent to the corresponding calibration reference.

Figure 2: Example plate layout with controls.



QC samples

QC samples provide a quality control check to determine if the plate performed as expected. These control samples are either pooled human plasma or serum, depending on the sample types being run. Plate quality is assessed by comparing the median of these samples per SOMAmer Reagent to the corresponding QC reference.

SOMAmer Reagent controls

SOMAmer Reagent controls are incorporated into each individual sample at known concentrations before hybridization and enable computing sample-specific scale factors to normalize samples and assess quality.

Data normalization steps

Normalization of raw proteomics counts involves multiple steps to remove systematic biases across samples and plates that may arise during the Illumina Protein Prep workflow. These adjustments are performed across four nonconsecutive steps: hybridization normalization, external reference median normalization, calibration, and plate scaling (Figure 3).

Hybridization normalization

This step corrects for sample-to-sample variability that can occur during the hybridization and library preparation stages of the assay. Sample-specific scale factors are computed by comparing the counts of the hybridization normalization controls in each sample to their median values across all samples in the plate.

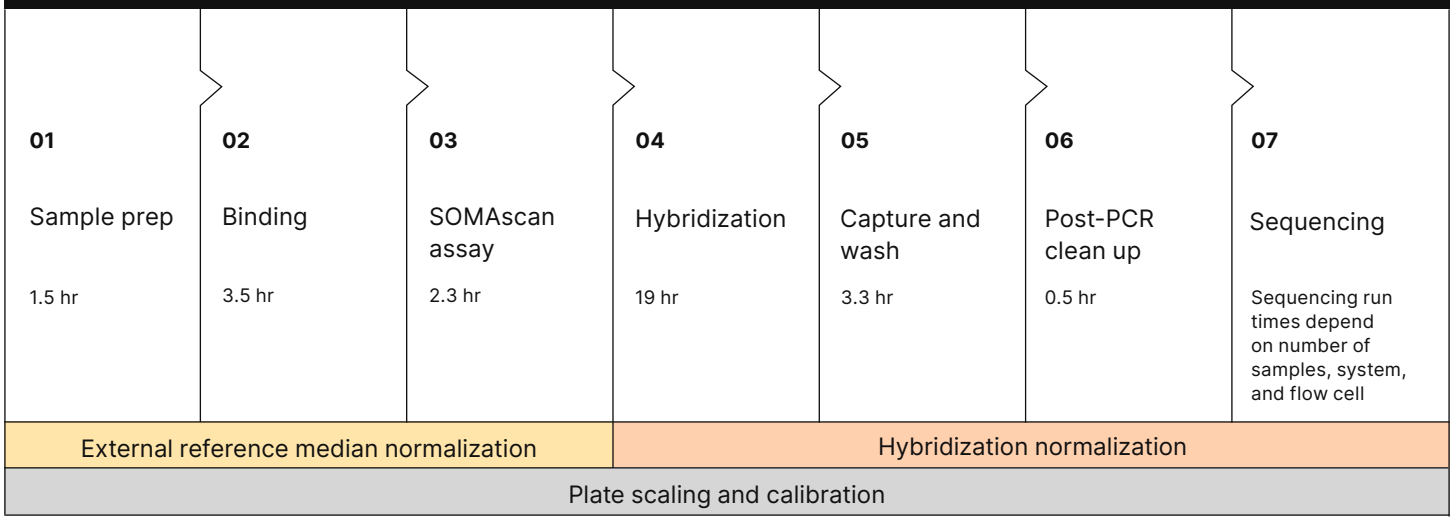
External reference median normalization

This step corrects for differences in the total protein abundance measurement of the non-control samples. A scale factor is calculated by comparing the observed protein measurements to a reference of expected values for each protein.

Calibration

Calibration is performed using the five calibrator samples outlined in the plate controls section and corrects for batch effects that impact individual SOMAmer Reagents. This step enables accurate comparisons between data produced from different plates and runs. For each SOMAmer Reagent, calibration corrects for variation between the calibrators and the expected reference values for the calibrator control.

Figure 3: Overview of normalization steps performed by the DRAGEN Protein Quantification pipeline.



Calibration is carried out in two steps:

- Step 1: Compares the calibrator medians to a reference derived from the sequencing system and adjusts each SOMAmer Reagent accordingly.
- Step 2: Compares the updated calibrator medians to a reference that is derived from sequencing systems, enabling comparisons between data generated on different instruments. The scale factor used to align the median of the calibrators to the reference value is applied to all samples on the plate.

Plate scaling

Plate scaling addresses variations in total protein measurements between plates by using the median of each SOMAmer Reagent measurement across calibrators and aligning them to an external calibration reference.

The scaling process occurs in two steps:

1. Step 1: Adjusts measurements based on a reference derived from the sequencing instrument, ie NovaSeq 6000 or NovaSeq X Systems.
2. Step 2: Refines the adjustment using a reference that remains across sequencing systems, enabling comparisons between data generated on different instruments. QC metrics from only the first scaling step are displayed in the DRAGEN Protein Quantification report.

QC check percent in tails

A QC step is performed at the end of normalization using the QC controls on each plate. This step compares the median of each SOMAmer Reagent measurement across the three QC sample replicates to an external QC reference to assign a SOMAmer Reagent–specific scale factor and a QC metric (QCCheckTailPercent).

The various normalization procedures implemented in the DRAGEN Protein Quantification pipeline minimize extraneous variability originating from distinct stages of the Illumina Protein Prep workflow (Figure 4).

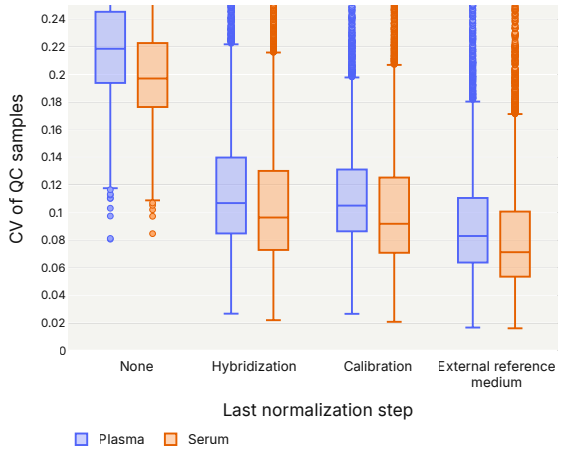
DRAGEN Protein Quantification output

The DRAGEN Protein Quantification pipeline outputs two types of files that are critical for analyzing Illumina Protein Prep data. The first is a series of ADAT files from each normalization step and the second is a DRAGEN QC report summarizing the QC of the run.

ADAT files

The DRAGEN Protein Quantification pipeline processes raw counts and generates a tab-delimited ASCII text file with an ADAT extension as the output for each normalization step performed. The ADAT file contains measurements for a series of analytes (columns) across a series of samples (rows) and includes analyte description and sample description information.

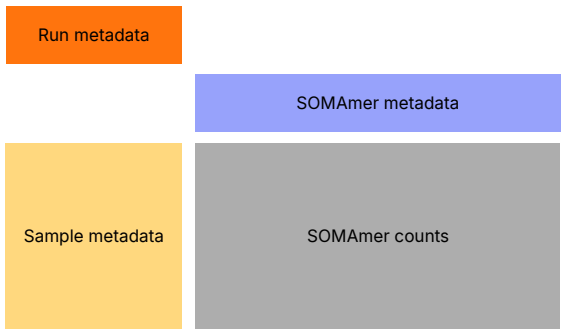
Figure 4: Impact of normalization on data variability.



Normalization of raw assay data results in reduced intraplate coefficient of variation (CV). Data were generated using the Illumina Protein Prep 6K assay with plasma and serum QC samples. The Illumina Protein Prep 9.5K assay with median CV ~5.5%[a] is now available.

a. Expected median CV calculated using healthy donor samples.

Figure 5: ADAT file structure.



The ADAT files are tab-delimited files with the following structure (Figure 5):

- **Run metadata** contains metrics that are relevant to the whole run or individual plates (eg, RunID and metrics associated with QC check percent in tails)
- **SOMAmer metadata** contains metrics that are relevant to each SOMAmer Reagent or protein target (eg, SOMAmer ID, and the specific gene and target IDs)
- **Sample metadata** contains metrics that are relevant to each sample (eg, Sample ID, Plate ID, and QC Metrics)
- **SOMAmer counts** are a matrix of NGS counts that reflect the relative abundance of each SOMAmer in each sample

The format is designed to provide flexibility for the number of samples as well as the number and types of analyte and sample descriptors and can be used for downstream tertiary analysis to gain biological insights.

DRAGEN QC Report

A DRAGEN Protein Quantification QC report accompanies the ADAT file and summarizes the QC of each individual sample and the entire plate. Individual sample QC is derived from the hybridization normalization and external reference median normalization steps. Plate QC is derived from the calibration normalization and QC steps.

Using ADAT files for tertiary analysis

The normalized ADAT file can be immediately used for further data analysis with Illumina Connected Multiomics via Illumina Connected Analytics. Illumina Connected Multiomics is equipped with a graphical user interface for data management and selection and can be used for advanced data interpretation, including differential protein expression, pathway analysis, and gene set enrichment analysis. The contents of the ADAT file can also be analyzed using open-source packages provided in R (SomadataIO) and Python (Canopy).

Summary

Normalization of raw proteomics data is essential to minimize assay and sample bias that may arise during the Illumina Protein Prep workflow. The DRAGEN Protein Quantification pipeline uses multiple normalization and calibration steps to reduce systematic variation while minimizing the risk of removing true biological variation, providing more accurate results. This pipeline is available on cloud or on-premises servers. On cloud, the DRAGEN Protein Quantification pipeline launches automatically following sequencing on the NovaSeq 6000 or NovaSeq X Systems, generating a DRAGEN QC report and ADAT files with normalized protein counts. The ADAT file output can then be readily used for downstream tertiary analysis with Illumina Connected Multiomics to enable deeper biological insights into the proteome.

Learn more →

Illumina Protein Prep

DRAGEN secondary analysis

Illumina Connected Analytics

1.800.809.4566 toll-free (US) | +1.858.202.4566 tel
techsupport@illumina.com | www.illumina.com

M-GL-03205 v1.0