Integrating array- and NGS-based proteomics data

Bridging SomaScan® Platform array-based proteomics outputs with Illumina Protein Prep NGS data enables unified analysis

Maximize insights by harmonizing data from multiplatform studies

Minimize platform-specific biases with highly correlated and concordant data

Streamline proteomics data analysis with Illumina Connected Multiomics

Introduction

Integrating protein abundance measurements across different assay technologies is an increasingly important requirement for multiplatform studies, longitudinal analyses, and maintaining data continuity. As study sizes increase and technologies evolve, there is a growing need for interoperability between existing data sets and newly generated ones across proteomic analysis platforms. The Illumina Protein Prep assay and the SomaLogic® SomaScan Platform both provide high-value proteomic information but differ materially in measurement chemistry, data structures, and normalization conventions. Combining data from these two platforms is possible with a clear bridging strategy that delivers rich biological insights, while minimizing platform-specific biases.

Both assays use SOMAmer® (slow off-rate modified aptamer) Reagents for protein identification but use fundamentally different molecular processes to generate proteomic data. Illumina Protein Prep uses next-generation sequencing (NGS) to quantify and detect proteins while the SomaScan Platform uses an array-based readout. Because both readouts are optimized for minimal background, they correlate well when comparing an individual analyte across multiple biological samples. However, due to differences between the two assays, including their units, dynamic range, and normalization references (Figure 1), the values and relative counts between analytes will differ between the readouts.

The goal of bridging between Illumina Protein Prep and SomaScan Platform data is to harmonize abundance measurements so that results can be compared more easily across platforms.

This technical note outlines the conceptual framework for bridging Illumina Protein Prep data with existing SomaScan Platform data sets using Illumina Connected Multiomics software. We provide a reproducible, transparent, and scientifically grounded approach that enables Illumina and collaborative teams to leverage existing SomaScan Platform data sets while incorporating new Illumina protein measurements into unified analyses.

Illumina Protein Prep data normalization

Array-based data can be bridged to NGS counts by applying the standard NGS data normalization pipeline directly to raw array counts. Data normalization corrects for intra- and inter-assay variability, standardizes data, and enables meaningful comparisons across samples, plates, assay runs, and platforms. The data normalization process consists of three sequential steps that correct for technical variability that can be introduced during hybridization, NGS library preparation, or liquid-handling portions of the assay (Figure 1).

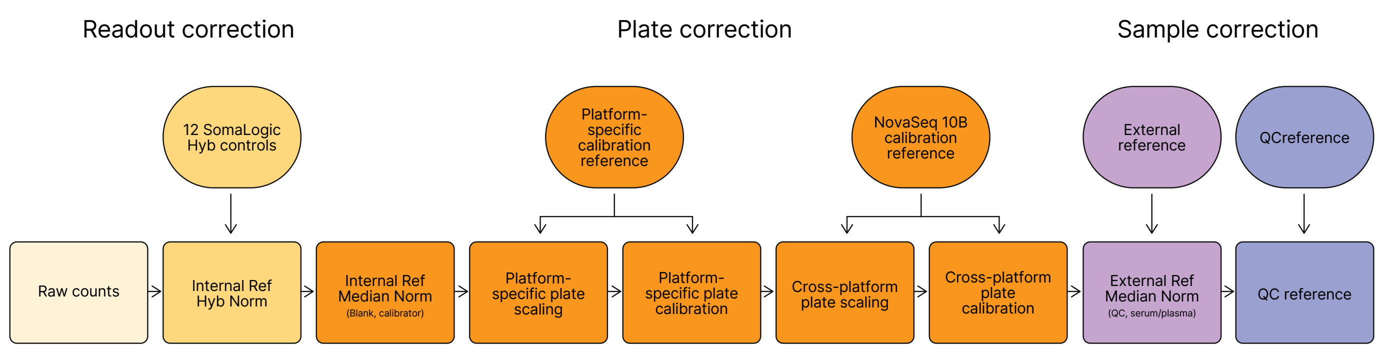

Figure 1: Sequential normalization steps in the Illumina Protein Prep analysis pipeline

Data normalization includes readout correction, plate correction, and sample correction steps to reduce variability that can be introduced during hybridization, library preparation, or liquid-handling portions of the assay. These steps are performed sequentially and an individual ADAT file is produced for each normalization step.

The normalization process includes the following steps:

- **Readout correction**: This step uses SOMAmer Reagent controls to reduce technical variability introduced in the NGS library preparation process
- **Plate correction**: This step uses calibrator controls as positive controls to correct for biases between plates; an external reference provided by Illumina enables comparison between Illumina Protein Prep plates from different assay runs
- **Sample correction**: This step uses protein abundance in each sample to reduce technical variability introduced during protein quantification and correct for differences in overall protein concentration

Bridging strategy

The purpose of bridging data sets is to place protein abundance measurements from relative fluorescence unit (RFU) values generated by the SomaScan Platform onto a scale that is interpreted alongside the NGS-based Illumina Protein Prep assay (Figure 2). Bridging can be applied during retrospective data integration, ongoing cohort expansion, and studies in which Illumina Protein Prep and SomaScan assays are used in parallel or sequentially.

Illumina Protein Prep 9.5K and SomaScan 11K assay both use identical calibrator controls for plate correction. As a result, when the cross-platform NGS calibration reference is applied to array data in Illumina Connected Multiomics, it will bridge that data to the NGS scale, making data from these two assays comparable. The assessment of the array performance is preserved in this normalization via the platform-specific calibration step in which the calibrators are compared to an array reference and a quality control (QC) metric is calculated.

Bridging provides highly correlated NGS and array data

To assess this bridging strategy, paired pooled plasma samples were assayed on both Illumina Protein Prep 9.5K Plasma assay and the SomaScan 11K assay. The samples were assayed a second time on Illumina Protein Prep to generate a baseline for comparing the interplatform results (array to NGS) with intraplatform results (NGS to NGS). NGS data sets were normalized using Illumina Protein Prep normalization, while the array data set was normalized using array normalization before bridging and NGS normalization after bridging (Figure 2).

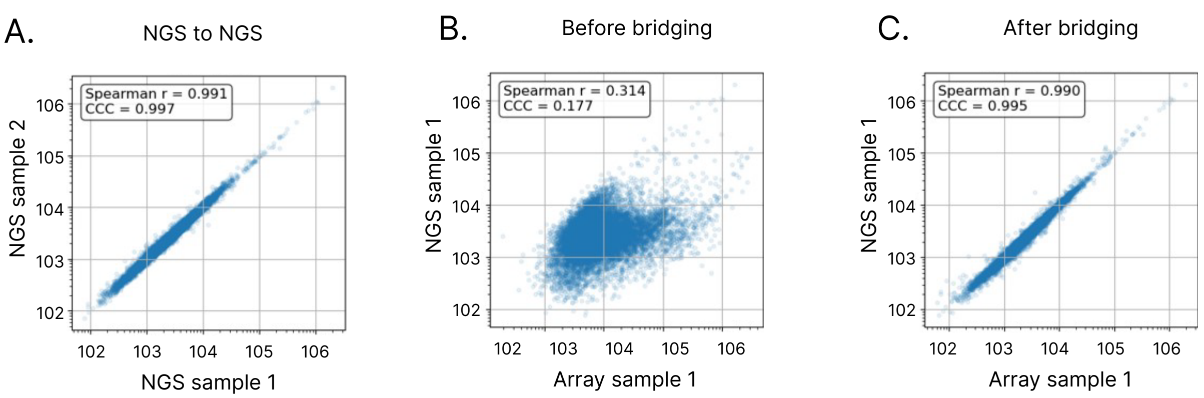

Figure 2: Spearman correlation coefficient between NGS- and array-based data before and after bridging for a single sample demonstrates data harmonization

Representative pooled plasma samples from different assay runs showing (A) excellent correlation between data from individual NGS runs (B) poor correlation between NGS and array data without bridging, primarily due to NGS specific dynamic range compression, and (C) improved correlation between NGS and array data, increasing to intraplatform levels with bridging.

Bridging aligns differences between NGS- and array-based readouts to yield comparable data (Figure 2). An important aspect of this correction is the relative scaling between SOMAmer Reagents, which differs significantly between NGS and array due to an additional dynamic range compression step (DRC) in the NGS readout that reduces sequencing variability across SOMAmer Reagents. This difference is evident even within a single pooled plasma sample, where intraplatform variance (Figure 2A) is markedly lower than interplatform variance (Figure 2B) before bridging. However, bridging overcomes this readout difference and aligns the array data to NGS count at a scale comparable to intraplatform variance (Figure 2C), quantified by the Spearman correlation coefficient and Lin's concordance coefficient. This quantitative similarity between intraplatform and interplatform variance after bridging is seen across the full distribution of pooled plasma sample comparisons (Figure 3).

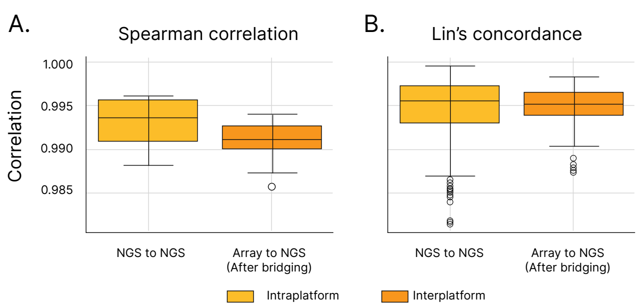

Figure 3: Distribution of Spearman correlation and Lin's concordance correlation coefficients across samples shows consistent effect of bridging in aligning data

Distribution of (A) correlation and (B) concordance across 18 pooled plasma comparisons between NGS- and array-based platforms. Correlations and concordance of array-to-NGS readouts after bridging across many pooled plasma samples is comparable to NGS-to-NGS correlations.

Dynamic range compression (DRC) is used to even out the SOMAmer Reagent concentrations from a 5-log to a 2-log dynamic range. Without DRC, the most abundant SOMAmer Reagents would occupy the majority of sequencing readout capacity. Each sample would require extremely high sequencing depth to cover the unabundant SOMAmer Reagents and accurately quantify them. With DRC, probes are grouped based on their observed abundances under no DRC condition and each group is compressed by a different set ratio.

Bridging improves count concordance between NGS and array data across SOMAmer panel

As part of the same study, a cohort of 32 paired plasma samples were assayed using the Illumina Protein Prep 9.5K, Plasma assay and the SomaScan 11K assay on two different Tecan liquid-handling setups. The samples were assayed a second time on the Illumina Protein Prep 9.5K assay to generate a baseline for intraplatform results (NGS to NGS). NGS data sets were normalized using Illumina Protein Prep normalization, while the array-based data sets were normalized using both array normalization (before bridging) and NGS normalization (after bridging). Intraplatform correlations with NGS using the two setups and interplatform correlations between NGS and array outputs were assessed (Figure 4, Figure 5).

Because both readouts measure relative SOMAmer Reagent concentration, individual SOMAmer Reagent counts across the biological cohort generally correlate. That is, the data moves in the same direction between array and NGS readouts, even without bridging and is quantified using the Spearman correlation coefficient (Figure 4, Figure 5A). However, SOMAmer Reagent count concordance, which is a measure of how close the actual counts are, is significantly lower in interplatform (array to NGS) comparisons than in intraplatform (NGS to NGS) comparisons. This is because of differences between the readout methodologies, including units, dynamic range, and normalization references. Bridging aligns the array data to the NGS readout scale, enabling comparison between counts per SOMAmer Reagent per sample. This is quantified by Lin's concordance correlation coefficient, which increases to intraplatform levels with bridging (Figure 5B).

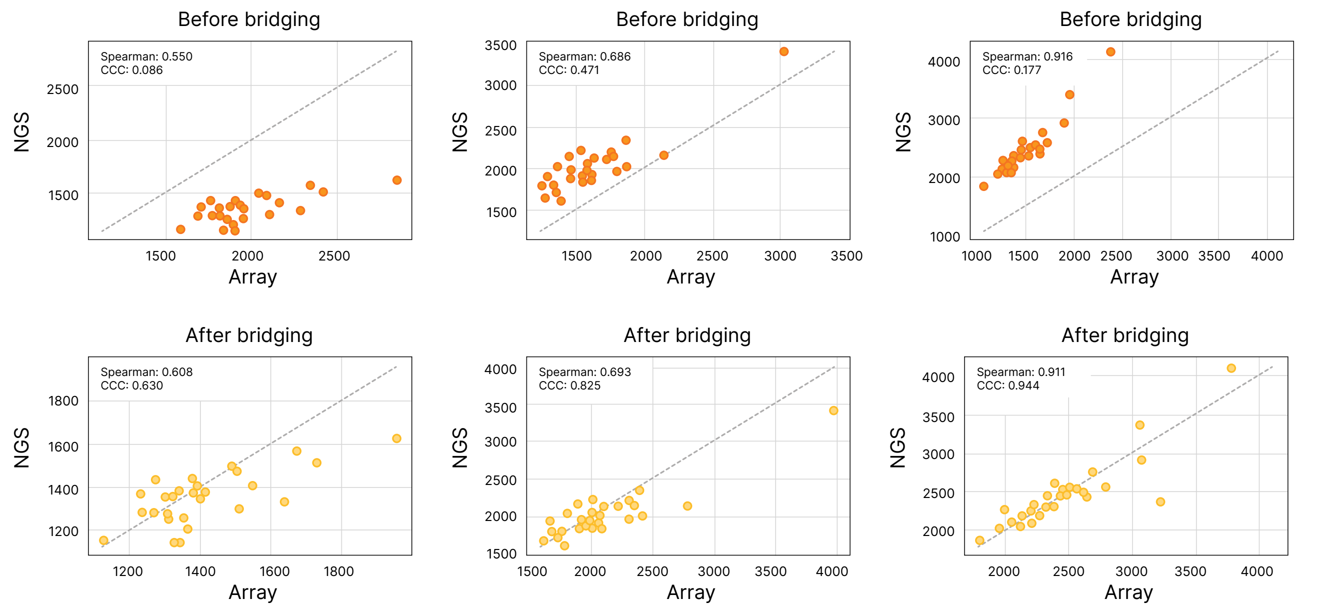

Figure 4: Examples of array to NGS correlations for three SOMAmer Reagents before and after bridging

Bridging maintains correlation and increases concordance between NGS and array counts for each SOMAmer Reagent across a biological cohort.

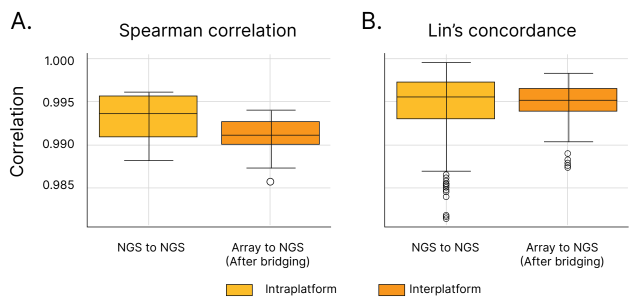

Figure 5: Distribution of Spearman correlation and Lin's concordance correlation coefficients across all human SOMAmer Reagents demonstrates consistent concordance with bridging

(A) Distribution of intraplatform NGS-to-NGS correlations across all SOMAmer Reagents is comparable with NGS-to-array correlations. (B) With bridging, the distribution of interplatform concordance approaches that of intraplatform concordance.

Summary

This technical note demonstrates that assay bridging provides highly correlated and concordant data from NGS- and array-based platforms. An effective bridging strategy allows labs to extend previously generated data from the SomaScan Platform with new Illumina Protein Prep measurements, enabling multiplatform biomarker discovery, supporting assay transition planning, and ensuring continuity of analyses across evolving proteomic pipelines.

Learn more →

Illumina Protein Prep

Achieving accurate and reliable protein detection with Illumina Protein Prep

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

M-GL-04147 v1.0