








Synthetic plasma pool cohort correction for affinity-based proteomics datasets allows multiple study comparison

Dries Heylen ^{1,2,*}, Murih Pusparum ^{2,3}, Jurgis Kuliesius⁴, Jim Wilson^{4,5}, Young-Chan Park⁶, Jacek Jamiołkowski⁷, Valentino D'Onofrio ⁸, Dirk Valkenburg ³, Jan Aerts ⁹, Gökhan Ertaylan ², Jef Hooyberghs ¹

¹Data Science Institute, Theory Lab, Hasselt University, 3590 Diepenbeek, Belgium

²Flemish Institute for Technological Research (VITO), Mol, Belgium

³Hasselt University, Data Science Institute, 3590 Diepenbeek, Belgium

⁴Centre for Global Health Research, University of Edinburgh, Edinburgh BioQuarter, Edinburgh EH16 4UX, United Kingdom

⁵MRC Human Genetics Unit, University of Edinburgh, Western General Hospital, Edinburgh, EH4 2XU, United Kingdom

⁶Institute of Translational Genomics, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany

⁷Department of Population Medicine and Lifestyle Diseases Prevention, Medical University of Białystok, 15-089 Białystok, Poland

⁸Center for Vaccinology, Ghent University and Ghent University Hospital, 9000 Ghent, Belgium

⁹Augmented Intelligence for Data Analytics (AIDA) Lab Department of Biosystems KU Leuven, Leuven, Belgium

*Corresponding author: Dries Heylen, E-mail: dries.heylen@uhasselt.be

Abstract

Proteomics stands as the crucial link between genomics and human diseases. Quantitative proteomics provides detailed insights into protein levels, enabling differentiation between distinct phenotypes. OLINK, a biotechnology company from Uppsala, Sweden, offers a targeted, affinity-based protein measurement method called Target 96, which has become prominent in the field of proteomics. The SCALLOP consortium, for instance, contains data from over 70,000 individuals across 45 independent cohort studies, all sampled by OLINK. However, when independent cohorts want to collaborate and quantitatively compare their target 96 protein values, it is currently advised to include 'identical biological bridging' samples in each sampling run to perform a reference sample normalization, correcting technical variations across measurements. Such a 'biological bridging sample' approach requires each of the involved cohorts to resend their biological bridging samples to OLINK to run them all together, which is logistically challenging, costly and time-consuming. Hence alternatives are searched and an evaluation of the current state of the art exposes the need for a more robust method that allows all OLINK Target 96 studies to compare proteomics data accurately and cost-efficiently. To meet these goals we developed the Synthetic Plasma Pool Cohort Correction, the 'SPOC correction' approach, based on the use of an OLINK-composed synthetic plasma sample. The method can easily be implemented in a federated data-sharing context which is illustrated on a sepsis use case.

Keywords: proteomics; biomarkers; normalization; protein quantification

Introduction

Proteins are the product of gene expression and are considered the building blocks of life by mediating the biochemical activities of cells and tissues [1]. The field of proteomics has seen remarkable advancements, with a stream of >70 000 publications since 2010, reflecting its significant impact on biomedical research [2]. Advances in high-throughput proteomics have lowered costs and expanded its application, allowing researchers to uncover associations among genes, proteins, and phenotypes. In clinical practice, proteomics promises improvements in early diagnosis, treatment planning, and health monitoring, ultimately improving patient outcomes.

Proteomics is expected to play an important role in precision medicine, though achieving its potential in clinical practice requires both clinical validation [3] and large, interoperable

datasets from multiple cohorts [1]. It is therefore crucial to have effective integration of existing and future proteomics data.

One widely adopted technology for targeted proteomics is developed by OLINK, a biotechnology company from Uppsala, Sweden, which employs a targeted, affinity-based approach. The proximity extension assay (PEA) quantifies the abundance of specific, preselected proteins within a biological sample, typically blood. Unlike mass spectrometry, which detects proteins by mass-to-charge ratio without prior knowledge on the protein target, OLINK's PEA employs antibodies linked with complementary oligonucleotides that hybridize and extend by using a DNA polymerase upon binding to a target protein [4, 5]. The initial concentration of the protein target is measured by the concentration of the generated DNA amplicon using quantitative PCR (qPCR) [6]. This method can currently measure up to

Received: August 30, 2024. Revised: November 5, 2024. Accepted: December 3, 2024

© The Author(s) 2024. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

For commercial re-use, please contact journals.permissions@oup.com

1100 different proteins, grouped into 15 protein panels related to cardiometabolic disorders, cell regulation, cardiovascular diseases, immune system, oncology, inflammation, metabolism, and neurology.

OLINK expresses protein abundance using Normalized Protein eXpression (NPX) values, which are logarithmically related to protein concentration in the sample. After internal corrections via technical controls (see subsection ‘Experimental layout and data normalization: IPC controls’), the NPX value is essentially equivalent to the negative qPCR ct-value (threshold cycle) thus following a log₂ scale.

Collaborative studies with large sample sizes advance the application of proteomics in precision medicine from concept to clinical application. A challenge, however, is addressing systematic technical variation across samples to enable accurate quantitative comparisons of protein expression values, both within and across cohorts. Since cohorts may vary in demographics or disease status, such normalization helps ensure that observed differences reflect true biological variation. Current methods for removing unwanted technical variation in proteomics, like ComBat and CONSTANd used in MS-based proteomics [7, 8], are not suitable for PEA, as they rely on mass spectrometry-specific assumptions. In affinity-based proteomics, quantile and median signal normalization are commonly used [9], but these assume similar data distributions or medians across samples. This approach is unsuitable for pathology data, where disease and control samples often exhibit different protein expression profiles.

OLINK’s recommended PEA normalization for intra-cohort analysis involves ‘biological bridging samples’ (see section Materials and Methods), where a minimum of eight distinct samples are included in each study plate. This approach is essential for robust study design but presents logistical challenges for combining cohort studies, as cohorts typically use different biological bridging samples. OLINK also recommends using identical biological bridging samples for data normalization in inter-cohort comparisons. This approach requires collaborating cohorts to rerun their biological bridging samples together in a single quantification run to establish a new reference set. This process is expensive, time-consuming, logistically challenging, and depends on the availability and quality of the original biological bridging samples for each cohort. To circumvent these additional costs and efforts, consortia seek alternative methods that do not require rerunning biological bridging samples across cohorts.

To address these challenges, we introduce a new normalization method for PEA, the Synthetic Plasma Pool Cohort Correction (SPOC Correction). The Materials and Methods section provides details on the cohorts and data used, as well as the current state-of-the-art OLINK normalization practices. In the Results and discussion, we first show why the current alternative normalization methods fall short before introducing our SPOC Correction, demonstrating its effectiveness and universality across diverse cohorts. To illustrate its applied value in a federated data sharing context, we present a case study on sepsis. Finally, we conclude by providing the algorithm freely on GitHub to facilitate its adoption.

Materials and methods

Description of cohorts

This subsection describes the set-up of several cohort studies that were used as the basis for the analysis in this manuscript.

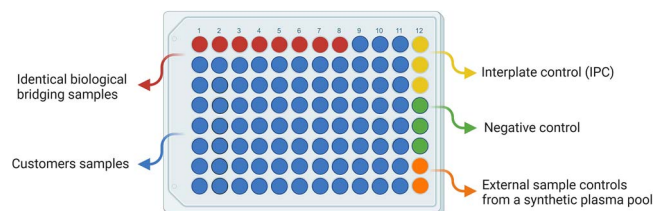


Figure 1. Schematic overview of OLINK’s sampling procedure. The well plate configuration is illustrative, and in practice the sample positions can be random.

The I AM Frontier (IAF) cohort was set up to support the development of a precision health-driven proof-of-concept aimed at advancing towards personalized prevention and health promotion. The cohort ran for 12 months as a longitudinal small-scale cohort study ($n=30$) in the Antwerp region of Flanders, Belgium. Participants were recruited as undiagnosed employees of the research institute hosting the study; they did not have a clinical diagnosis and were between 45 and 60 years old. The IAF sample collection started in March 2019 and ran for 12 months. The cohort contains OLINK proteomics data on a bimonthly basis, all OLINK protein panels were included (see Table 1). Full access to this data was obtained for use in this work.

A prospective cohort study hereafter referred to as the Fast Assay for Pathogen Identification and Characterization (FAPIC) cohort was set up to identify biomarkers of inflammation for the prognosis and diagnosis of bloodstream infection and sepsis. In the FAPIC cohort, samples were collected from 406 suspected sepsis episodes in the emergency room of the Jessa hospital at the time of admission (Hasselt, Belgium) [10]. The OLINK protein inflammation panel was included (see Table 1). Full access to this data was obtained for use in this work.

The Systematic and Combined Analysis of Olink Proteins (SCALLOP) consortium is a collaborative framework for discovery and follow-up of genetic associations with proteins on the OLINK proteomics platform [11]. Data was collected for >70 000 patients and controls across 45 independent cohort studies. A varying selection of OLINK protein panels was included. Federated data access was obtained for three studies involved in the SCALLOP consortium (Supplementary Table 1). For these studies, only data from the synthetic plasma pool composed by OLINK (see experimental layout and data normalization) was used and no individual specific biological data was accessed.

Experimental layout and data normalization

The experimental layout of OLINK measurements follows the 96-well plate format (Fig. 1). Each plate consists of a combination of biological samples and control samples. All the different sample types are described below. The work in this paper mainly focuses on the use of external sample controls from a synthetic plasma pool as alternative for the identical biological bridging samples.

Inter-plate Control (IPC) are included in triplicate on each plate and these are run as normal samples. The IPC are a pool of 92 antibodies, each with one pair of unique DNA tags positioned in fixed proximity and can be seen as a synthetic sample, expected to give a high signal for all proteins. The median of the IPC triplicates is used to normalize each protein to compensate for potential variation between plates [12].

Negative controls are also included in triplicate on each plate and consist of buffer run as a normal sample. These are used to monitor any background noise generated when DNA tags come in close proximity without prior binding to the appropriate protein.

Table 1. Sample overview (I-AM frontier & FAPIC).

Sample overview	IAF	FAPIC
Total number of samples	210	400
Longitudinal sample distribution	7 x 30 samples	250 samples batch 1 150 samples batch 2
Number of batches sent for sampling to OLINK	4	2
Number of 96 well plates required in total	4	6
OLINK controls per plate (fixed number for all cohorts)	2	2
Total number of synthetic plasma controls from OLINK	8	12
Total number of measured proteins	1070 (All Panels)	92 (Inflammation panel)

The negative controls set the background levels for each protein assay and are used to calculate the limit of detection [12].

External sample controls are used to assess potential variation between and within plates, through the calculation of inter-protein and intra-protein coefficients of variability (CV) [10]. Samples originate from a synthetic plasma pool composed by OLINK. Because only a small volume per sample is required (1 μ l per sample), one pool provides enough material for the service provider to execute many studies over the course of several years. A pool renewal took place at OLINK in October 2019.

Identical biological bridging samples are a minimum of eight biological samples included by the customer in all plates of a study. These samples are used as cornerstone for the default OLINK reference sample normalization described above. Each involved plate includes these eight distinct samples (Fig. 3A), ensuring matching across different plates. These identical biological bridging samples are referred to as bridging reference samples in the reference sample normalization protocol [12].

Default reference sample normalization with identical biological bridging samples

To perform data normalization between studies sampled by OLINK at different time points, or between different batches from a single (longitudinal) study sampled by OLINK at different time points, OLINK does not recommend relying solely on *Inter-plate Control*. For study designs which are not a priori randomized, which is the case we focus on, technical variability is minimized by running *Identical biological bridging samples* on each plate. A default reference sample normalization as imposed by OLINK allows comparison within one cohort as long as each plate of a cohort includes 8 to 16 of their samples as such identical biological bridging samples. Reference sample normalization is then performed in the following way [12]:

1. Choose a reference plate to normalize towards.
2. For each protein and plate, calculate the pairwise difference in protein expression, i.e., the NPX value, for each of the overlapping samples with the reference plate.
3. Estimate the plate- and protein-specific normalization term by calculating the median for the pairwise differences calculated in step 2.
4. For each protein and plate, add the plate- and protein-specific normalization term from step 3 to each value, to normalize it to the reference plate chosen in step 1.

Results and discussion

Evaluation of alternative PEA normalization

To express the abundance of a protein Olink uses the Normalized Protein eXpression (NPX), as explained in the introduction. We

will use the letter X throughout the paper for the NPX value and introduce the following notation:

$$TX_{p,w}(i) = \text{he NPX value of protein nr } i \text{ in well nr } w \text{ of plate nr } p.$$

Note that in the contexts of cohorts each well corresponds to the sample of a subject.

OLINK's reference sample normalization (see Materials and Methods) allows the quantitative comparison of protein NPX values for samples within one cohort that were sampled on different 96-well plates, at different moments in time (i.e., different batches). As long as all plates of a cohort used 8 to 16 of their samples as biological bridging samples, reference sample normalization can be performed to address the technical variation induced by the measurement process. On the other hand, cohorts sampled independently by OLINK have to resample their biological bridging samples along with those of potential collaborators before comparing their data across cohorts, which is a costly and time-consuming process.

Therefore some SCALLOP consortium members (see Materials and Methods) previously used a fractional rank normalization method as the alternative in the field to work around this default reference sample normalization procedure [11]. With the rank normalization method, the NPX protein values were rank-based, inverse normal transformed, and standardized to the unit variance. This rank based Inverse Normal Transformation (INT) is presented in equation (1) where the plate nr p is a fixed value:

$$INT(X_{p,w}(i)) = \Phi^{-1} \left\{ \frac{\text{rank}_w(X_{p,w}(i)) - c}{n + 1 - 2c} \right\}, c \in [0, 1/2] \quad (1)$$

$X_{p,w}(i)$ is the continuous NPX abundance measurement of protein i in well w . Each well corresponds to one of the n subjects, and rank_w is the subject rank of protein i when the NPX abundance measurements are placed in ascending order over the subjects. Here Φ^{-1} is the probit function, and $c \in [0, 1/2]$ is an adjustable offset. By default, the Blom offset of $c = 3/8$ is adopted.

This method aims to avoid OLINK technical variation between cohorts based on the principle that even though the measured proteins' absolute value might vary, the ratio between two proteins would not differ when measuring the same aliquot in different batches. To check the validity of these assumptions and to evaluate this principle, we used the I AM Frontier (IAF) cohort (see Materials and methods), which consists of healthy participants with bimonthly proteomics measurements over one year timespan. The participants were healthy and not expected to experience major health events during the study period. Therefore, if the fractional rank normalization procedure is appropriate for this type of data, we would expect minimal variation in fractional ranks of protein measurements across timepoints. Figure 2 shows that a rank fraction switch occurred often in the IAF cohort, indicating high within-person variability over time.

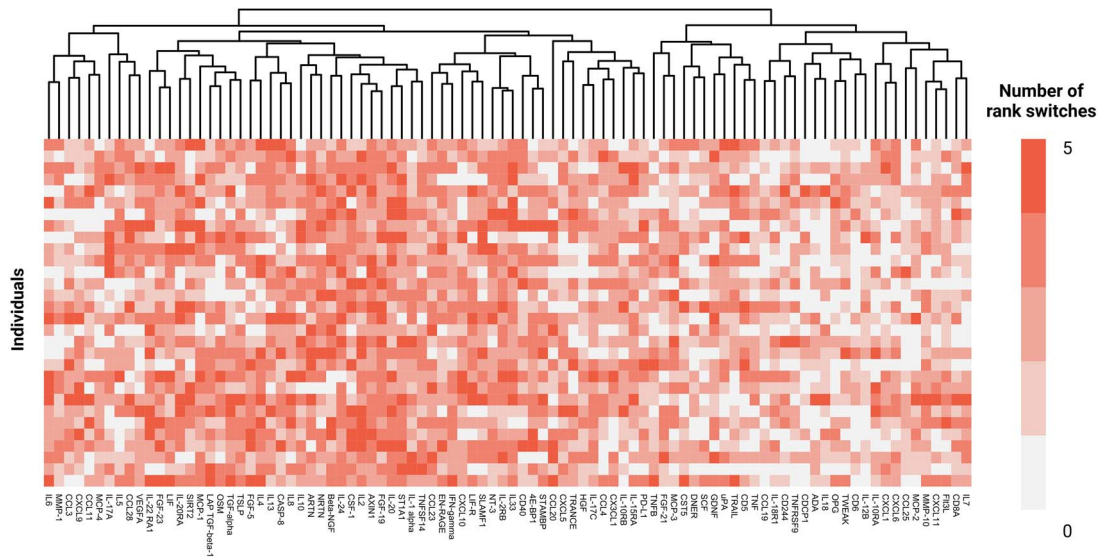


Figure 2. Evaluation of the rank-based, inverse normalized transformation (INT) applied to IAF cohort data. Rank switches throughout timepoints of measurement are shown for OLINK inflammation panel data for all 30 IAF individuals. Six consecutive protein collection periods are used (5 transitions). Ranks for each protein are assigned by ordering each individual's protein NPX value from high to low. Each individual's rank for each protein is evaluated against its rank from the previous time point. A rank switch for a protein is considered present if a person is at least 10% higher or lower up the ranking when comparing the position of an individual over two consecutive months. Hierarchical clustering is applied to the rank switch data to determine the order of the columns (proteins).

We are aware that this highly frequent longitudinal set-up is not the same as measuring identical aliquots repeatedly over time. However, these findings show the highly sensitive nature of these protein NPX measurements. Since protein values for many different people are situated in the same narrow range, applying the rank-based INT normalization technique might risk wrongly ranking individuals into specific fractions. These findings encourage the need for another cost-efficient, robust, computational normalization.

Therefore, the computational SPOC correction procedure (visualized in Fig. 3B) is proposed in this paper to make universal collaboration easier across OLINK cohorts.

SPOC correction for cohorts with shared external control samples

To aim for universality, the SPOC correction makes use of external sample controls consisting of a synthetic plasma pool (Fig. 1). The synthetic plasma pool is inserted by OLINK on every plate. It is a plasma pool that is used by OLINK to assess potential variation between and within plates through the calculation of inter- and intra-protein CV's.

The formulas below describe how the protein NPX values X are corrected by a term δ where plate a is used as the reference plate to which each other plate b is normalized. Equation (2) shows the reference sample normalization (corresponding to the four-step algorithm of Materials and Methods) and is denoted with the superscript ref . As normalization samples it uses the wells $k=1 \dots 8$ which contain the 8 biological bridging samples (Fig. 1). Each protein i in well (subject) w is consequently normalized as:

$$X_{p=b,w}^{ref}(i) = X_{p=b,w}(i) + \delta_{p=b}^{ref}(i) \quad (2)$$

$$\delta_{p=b}^{ref}(i) = \text{median}_{k=1:8} [X_{p=a,k}(i) - X_{p=b,k}(i)].$$

Note that the correction term δ is dependent on plate and protein, but independent of the well.

Equation (3) shows the SPOC correction, denoted with superscript $spoc$. Here the index k runs over the external sample controls in well 95 and 96 containing the synthetic plasma pool (Fig. 1).

$$X_{p=b,w}^{spoc}(i) = X_{p=b,w}(i) + \delta_{p=b}^{spoc}(i) \quad (3)$$

$$\delta_{p=b}^{spoc}(i) = \text{median}_{k=95,96} [X_{p=a,k}(i)] - \text{median}_{k=95,96} [X_{p=b,k}(i)]$$

Note that in the ref case we have eight different biological bridging samples which are repeated on both plates. This allows to consider them as eight coupled measurements, hence, use the median of the pairwise differences. In the SPOC case the external sample controls are not coupled across plates, hence we subtract the median of the two plates.

Fig. 4 compares the SPOC correction terms δ^{spoc} with the reference sample correction terms δ^{ref} for both the IAF cohort and the FAPIC cohort. Each dot represents a correction term δ for a specific protein. Reference sample correction terms were plotted on the y-axis and SPOC corrections were plotted on the X-axis. For the cohorts at hand, identical biological bridging samples are available and hence the reference sample correction is considered the golden standard. The R-squared value between the SPOC correction value and the reference sample correction value were significant (IAF cohort: 0.849 and FAPIC cohort: 0.882), showing the potential of the SPOC correction to be used as alternative for the reference sample normalization (Fig. 4A and B). There is no apparent bias between the two correction methods as the mean difference lines (in black) in the two Bland-Altman plots are close to zero with values of 0.06 and 0.03 (Fig. 4C and D).

SPOC correction for cohorts without shared external control samples

It is important to note that the external sample controls are from the same synthetic plasma pool in numerous cohorts, as one pool can be quantitatively sampled many times (only $1 \mu\text{l}$ per sample is used). However, a renewal of the synthetic plasma pool took

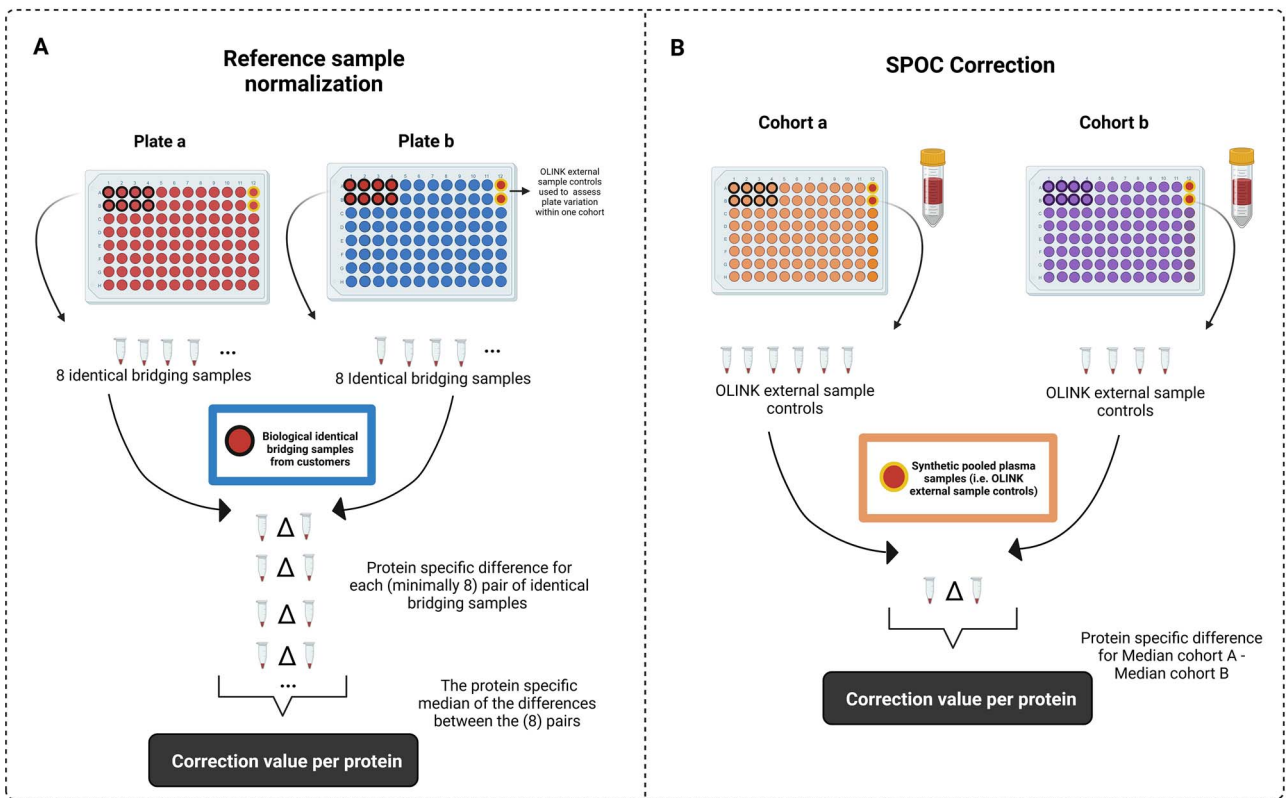


Figure 3. Default reference sample normalization versus SPOC correction procedure. (A) Default reference sample normalization of OLINK allows quantitative comparison of protein NPX values but requires cohorts to include at least 8 of their samples as biological bridging samples on each plate of the cohort. (B) The SPOC correction allows universal collaboration across OLINK cohorts using external sample controls consisting of a synthetic plasma pool.

place at OLINK in October 2019. This implies that cohorts whose external sample controls do not originate from the same synthetic plasma pool will require a plasma pool correction (Fig. 5). For the IAF cohort, OLINK proteomics data was sampled both before and after the OLINK pool renewal. Hence, both the old and the new plasma pool are available in the IAF study. This allowed the calculation of a 'plasma pool correction term'. The OLINK reference sample normalization was applied on the external sample controls from the old OLINK plasma pool, while taking the new plasma pool as a reference. This results in a pool correction value δ^{pool} that can be used to correct the plasma pool effect (Fig. 5).

Equation (4) shows how the SPOC correction can be applied for cohorts that do not share external sample controls from the same synthetic plasma pool. Note that the pool correction term neither depends on the plate or well, but is protein dependent.

$$X_{p=b,w}^{SPOC}(i) = X_{p=b,w}(i) + \delta_{p=b,w}^{SPOC}(i) + \delta^{pool}(i) \quad (4)$$

Within the IAF cohort, all proteins that OLINK offers using the Target 96 technique, were measured. This effort allows all Target 96 cohorts to apply the SPOC correction on all the protein panels available at OLINK, regardless of which of the two plasma pools is included. These values δ^{pool} are also included in the SPOC correction method that is made publicly available at <https://github.com/VITO-UHasselt-SPOC-correction/OLINK-Target-96-cohort-bridging>.

In Fig. 6 the synthetic plasma data of four studies involved in the SCALLOP consortium was assessed with a hierarchical clustering visualization. Figure 6A shows that the OLINK plasma pool renewal that took place in October 2019 clearly separates

cohorts in two groups based on which plasma pool that is used as external sample control data. A distinction is visible between cohorts sampled before October 2019, which contain the older synthetic plasma pool from OLINK (Helic Pomak and Orcades), and those sampled after October 2019, which contain the new synthetic plasma pool from OLINK (Bialystok Plus). Among the four displayed studies, also the I AM Frontier (IAF) cohort study was included. The IAF study has data sampled by OLINK at four distinct moments in time (see Materials and Methods). In the first batch, the old synthetic plasma pool was included by OLINK. In the last three batches the new plasma pool was included. Figure 6B shows the results of the internal reference sample normalization within the IAF study. This shows that the reference sample normalization, which performs the normalization based on biological bridging samples, corrects the protein NPX values of the new synthetic plasma pool towards the one of the old synthetic plasma pool.

Figure 6C on the other hand shows the result of applying the SPOC correction procedure within the IAF study. Similar to the reference sample normalization, the SPOC correction brings the corrected batches closer towards the first IAF batch. The demonstrated shift in sample similarity after correction that is comparable between Fig. 6B and Fig. 6C again confirms the validity of the SPOC correction strategy as an alternative to the reference sample normalization.

The SPOC correction relevancy in an applied setting

To show the potential impact and the practical use of the SPOC correction for quantitative comparison across cohorts we present

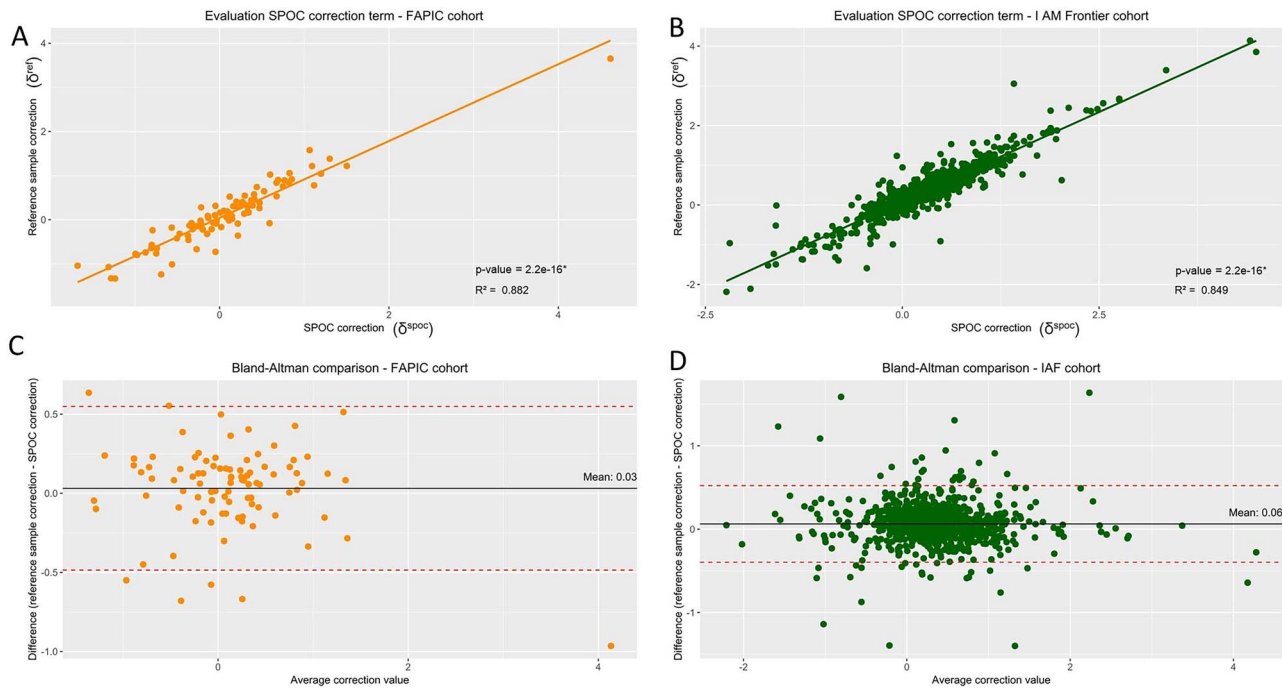


Figure 4. Comparison of the SPOC correction terms δ^{spoc} with the reference sample correction terms δ^{ref} . Each dot represents a correction value for a specific protein. (A-B) Correction values calculated with the reference sample correction method plotted on the y-axis against the correction values calculated with the SPOC correction method on the x-axis. (C-D) Bland-Altman plots to analyze the agreement between the two correction methods. The horizontal middle line indicates the average difference between the SPOC correction term and the reference sample correction term. Upper and lower 95% confidence intervals are indicated by the dotted lines. Left, for the inflammation protein panel from OLINK in the FAPIC cohort (92 proteins). Right, for all available protein panels from OLINK in the IAF cohort (1068 proteins). The samples that were run for IAF on the plates in batch 2 and batch 4 are used for the IAF plots. *with null-hypothesis H_0 = independent variables (i.e. SPOC correction terms) in the regression model explain the variability of the dependent variable (i.e. reference sample correction terms) in a random way. Based on this evaluation we consider the SPOC correction as a valid normalization that can be used when identical biological bridging samples are not present across different cohorts.

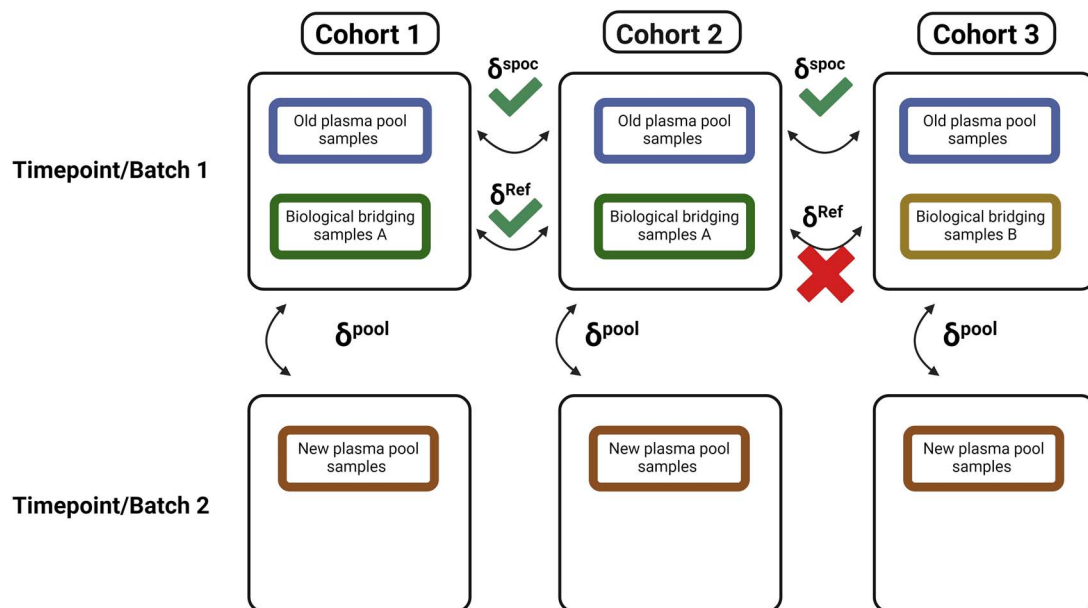


Figure 5. Landscape of OLINK proteomic studies sampled with a qPCR target 96 approach. A technical measurement variation can be bridged with a biological bridging sample correction (δ^{ref}) as long as these are available across all plates of the different cohorts or across sampling timepoints. If this is not the case (see cross sign) synthetic plasma samples correction (δ^{spoc}) can be used if their plasma sample is from the same synthetic plasma pool. If a different plasma pool is included a pool effect correction (δ^{pool}) is also needed.

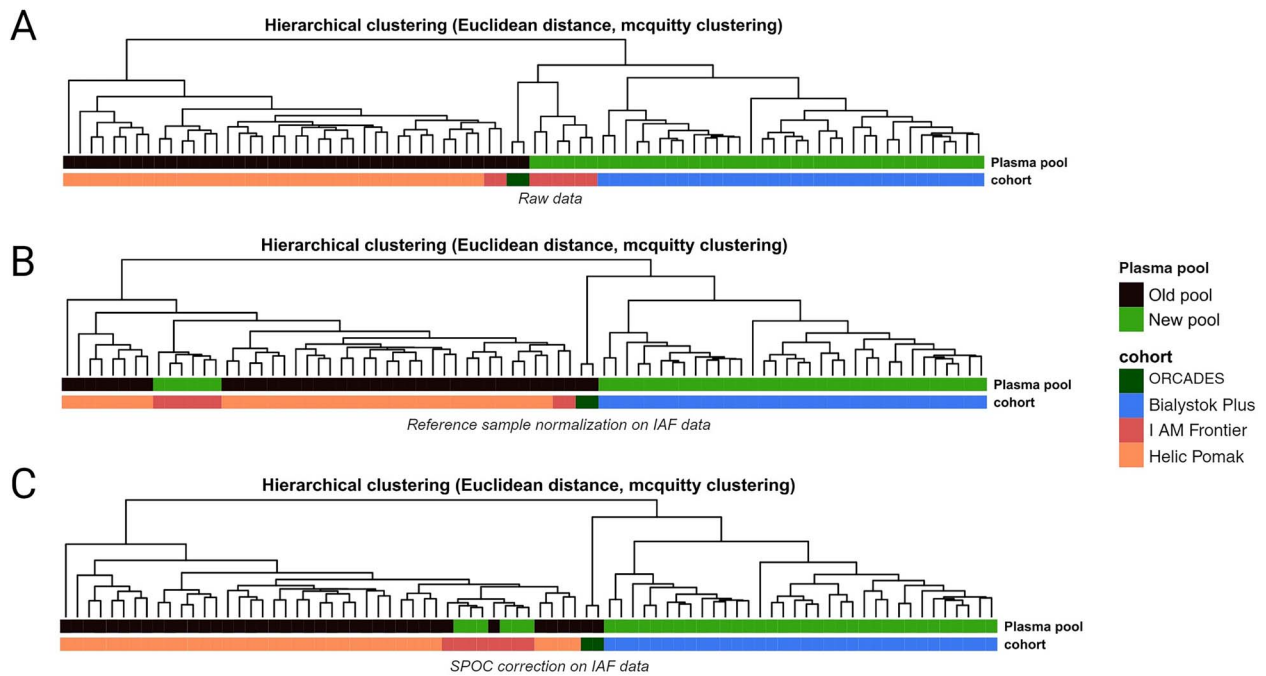


Figure 6. Unsupervised hierarchical clustering with a Euclidean distance measure and Mcquitty clustering method. Clustering is performed on protein NPX values for the target 96 metabolism panel of external sample controls. Two separate horizontal color coded bars indicate the cohort and plasma pool origin for each external sample control. (A) Clustering is shown for all external sample controls from four different cohorts involved in the SCALLOP consortium. Complete separation of the old versus new plasma pool is visible. (B) Reference sample correction applied on the IAF samples by computing the correction values with identical biological bridging samples. (C) SPOC correction applied on the IAF samples by computing the SPOC correction values with external sample controls. Complete clustering plots, including protein expression values, are attached as supplementary figures (Fig. S1-S3).

a use case which is fully implemented according to federated data-sharing principles. Figure 7 compares the healthy individuals from the IAF cohort with individuals with suspected sepsis from the FAPIC cohort.

To optimize the comparability across proteins and focus on the difference between both cohorts, the green reference intervals in Fig. 7 are centered with the median NPX values of the IAF cohort (i.e., the median IAF value for a protein is subtracted)

The IAF reference intervals are obtained through a non-parametric bootstrapping method (see section Code availability for a description of the detailed computation).

This analysis allows the identification of proteins that differ between cohorts in a case-control manner. In Fig. 7, proteins where the intervals do not overlap can be considered as potential markers for phenotype differences between healthy and suspected sepsis individuals. Several proteins shown on the right side of Fig. 7 (with the most significant margin between the green and orange intervals) have been previously reported in other studies to play a role in the pathology of sepsis, either by being increased (IL-6, IFN- γ , IL-17A, IL-8, MCP-1, MCP-4, IL20RA) or decreased (IL-5 and IL2) in patients with sepsis [13, 14]. For a protein as IL-24, our analyses suggest a potentially intriguing hypothesis with a significant difference between cases and controls, despite this protein not being commonly reported as a sepsis biomarker. This case-control workflow across cohorts is only possible when cohorts can be bridged by an adequate normalization. The FAPIC cohort and the IAF cohort did not share biological bridging samples. Only the SPOC correction allowed a quantitative comparison across these two cohorts with a distinct phenotype.

All scripts are made publicly available and are free to use (see Code availability below). The repository contains scripts

that run the SPOC correction as well as a script to establish the appropriate federated data-sharing setup. The bootstrapping steps are included, providing users with the intervals of their collaborating cohort, as displayed in Fig. 7. This enables users to perform a swift, quantitative comparison of their data to an external cohort of interest, in a federated way. The external sample control values for the IAF cohort are also available in the online repository so that users can opt to use these as 'baseline' reference pool.

This paper aims to enhance the utility of OLINK Target 96 cohort studies by enabling data comparisons across different cohorts. We would like to emphasize that this work complements recent initiatives within the SCALLOP consortium, which is exploring the pooling of samples from various Target 96 studies to be sampled together using the Next Generation Sequencing (NGS) OLINK *explore* technique [15]. This pooled OLINK *explore* dataset of SCALLOP consortium samples could serve as a 'reference pool' for harmonizing data across different measurement techniques. When combined with the SPOC correction technique presented here, the aforementioned SCALLOP consortium efforts could create a more consistent and interoperable data environment, aligning Target 96 proteomics studies with both current and future OLINK proteomics data.

In addition, the holistic approach of integrating proteomics with other omics platforms sheds light on the molecular transitions from genotype to phenotype. This integration holds the crucial potential to describe disease-related pathways, identify novel biomarkers for diagnostics and detect drug targets [16]. Such collective efforts and technological innovations in proteomics are not only advancing our understanding of the biology underlying health and disease but are also paving the way for the next

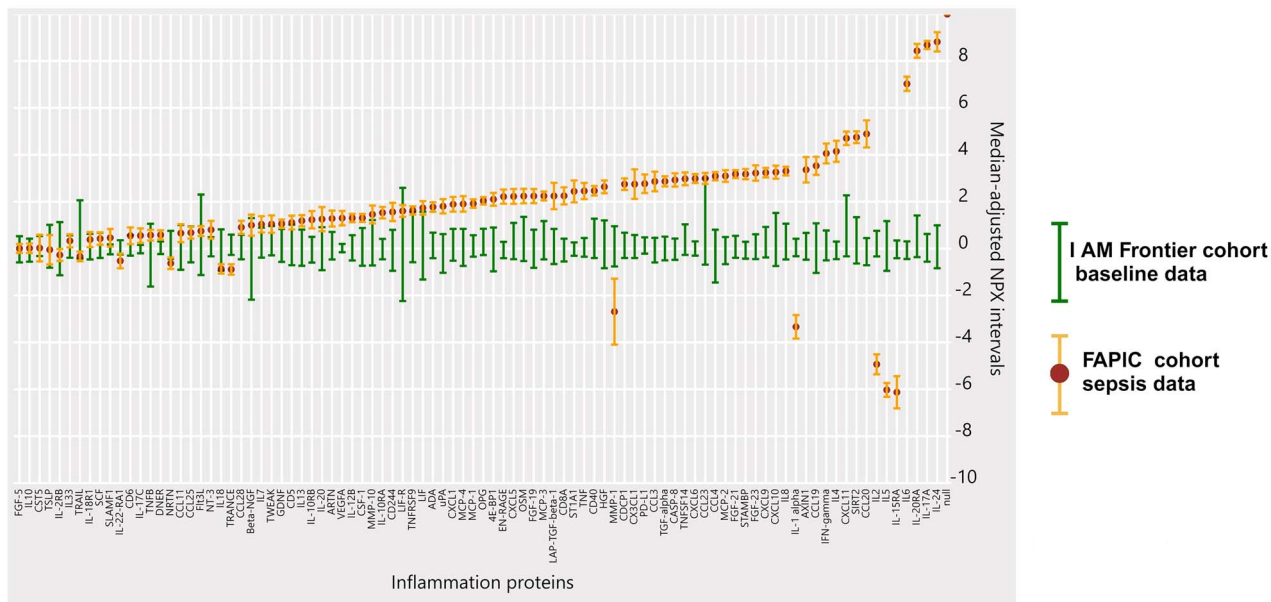


Figure 7. Federated transfer of protein ranges to compare protein NPX values between cohorts with distinct phenotypes (IAF-FAPIC). Green error bars show healthy NPX reference intervals based on the IAF data. The red data points show the difference in protein NPX value between the median of the IAF cohort and the median of the FAPIC cohort. The orange error bars represent a technical measurement error margin. To optimize the comparability across proteins and focus on the difference between both cohorts, the green reference intervals are centered with the median NPX value's of the IAF cohort. Intervals that do not overlap indicate a significant difference between the two phenotypes for the relevant protein.

generation of medical breakthroughs in disease management and therapy.

Key Points

- The SPOC correction approach allows qPCR target 96 sampled datasets to quantitatively compare their protein NPX values with each other by means of a cost-efficient computational method, leveraging research in the field of proteomics.
- As different cohorts might focus on different phenotypes and not all cohorts have a case-control set-up, the proposed correction methodology adds value to many studies and increases insights by obtaining differential protein profiles between distinct phenotypes.
- For cohorts that establish a (federated) data collaboration subsequent to their sampling procedure the SPOC correction provides a fast a cost-efficient alternative to the reference sample correction.

Acknowledgements

The FAPIC cohort was set up by the Department of Infectious Diseases and Immunity, Jessa Hospital, 3500, Hasselt, Belgium in collaboration with the Faculty of Medicine and Life Sciences, UHasselt, LCRC, Martelarenlaan 42, 3500, Hasselt, Belgium.

The Orkney Complex Disease Study (ORCADES) was supported by the Chief Scientist Office of the Scottish Government (CZB/4/276, CZB/4/710), a Royal Society URF to J.F.W., the MRC Human Genetics Unit quinquennial programme 'QTL in Health and Disease', Arthritis Research UK and the European Union framework program 6 EUROSPAN project (contract no. LSHG-CT-2006-018947). DNA extractions were performed at the Edinburgh Clinical Research Facility, University of Edinburgh. We would like to

acknowledge the invaluable contributions of the research nurses in Orkney, the administrative team in Edinburgh and the people of Orkney. VITO NV is a leading European research and technology organization that focuses on sustainable development. UHasselt Data Science Institute, focuses on interdisciplinary research and education in data science. Both institutes are based in Belgium.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Funding

D.H. is funded through a Hasselt University BOF grants (BOF200-WB29) D and VITO NV (R-11362). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability

The data underlying this article will be shared on reasonable request to the corresponding author.

Code availability

Scripts that extract data from a locally saved folder and subsequently apply all analyses steps in a federated way are available for usage at: <https://github.com/VITO-UHasselt-SPOC-correction/OLINK-Target-96-cohort-bridging>.

References

1. Correa Rojo A, Heylen D, Aerts J. et al. Towards building a quantitative proteomics toolbox in precision medicine: a

- mini-review. *Front Physiol* 2021;**12**:1394. <https://doi.org/10.3389/fphys.2021.723510>.
2. Mesri M. Advances in proteomic technologies and its contribution to the field of cancer. *Adv Med* 2014;**2014**: 1–25. Available from: /pmc/articles/PMC4590950/. <https://doi.org/10.1155/2014/238045>.
 3. Liu X, Luo X, Jiang C. et al. Difficulties and challenges in the development of precision medicine. *Clin Genet* 2019;**95**:569–74. <https://doi.org/10.1111/cge.13511>.
 4. Proximity T, Assay E, Ab OP. et al. White Paper PEA – A High-Multiplex Immunoassay Technology with qPCR or NGS Readout. Olink proteomics AB. <https://olink.com/>.
 5. Wik L, Nordberg N, Broberg J. et al. Proximity extension assay in combination with next-generation sequencing for high-throughput proteome-wide analysis. *Mol Cell Proteomics* 2021;**20**:100168. <https://doi.org/10.1016/j.mcpro.2021.100168>.
 6. Assarsson E, Lundberg M, Holmquist G. et al. Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. *PloS One* 2014;**9**:e95192. <https://doi.org/10.1371/journal.pone.0095192>.
 7. Maes E, Hadiwikarta WW, Mertens I. et al. CONSTANd : a normalization method for isobaric labeled spectra by constrained optimization. *Mol Cell Proteomics* 2016;**15**:2779–90. <https://doi.org/10.1074/mcp.M115.056911>.
 8. Phua SX, Lim KP, Bin GWW. Perspectives for better batch effect correction in mass-spectrometry-based proteomics. *Comput Struct Biotechnol J* 2022;**20**:4369–75 Available from: <https://doi.org/10.1016/j.csbj.2022.08.022>.
 9. Brauer BL, Wiredu K, Gerber SA. et al. Evaluation of quantification and normalization strategies for phosphoprotein phosphatase affinity proteomics: application to breast cancer signaling. *J Proteome Res* 2023;**22**:47–61. <https://doi.org/10.1021/acs.jproteome.2c00465>.
 10. D'Onofrio V, Heylen D, Pusparum M. et al. A prospective observational cohort study to identify inflammatory biomarkers for the diagnosis and prognosis of patients with sepsis. *J Intensive Care* 2022[cited 2022 May 23];**10**:13–3. Available from: <https://doi.org/https://jintensivecare.biomedcentral.com/articles/10.1186/s40560-022-00602-x>.
 11. Folkersen L, Gustafsson S, Wang Q. et al. Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nat Metab* 2020;**2**:1135–48. <https://doi.org/10.1038/s42255-020-00287-2>.
 12. Proteomics O. Data normalization and standardization. *White Paper* 2022;1–3. Available from: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://7074596.fs1.hubspotusercontent-na1.net/hubfs/7074596/05-white-paper-for-web-site/1096-olink-data-normalization-white-paper.pdf.
 13. Barichello T, Generoso JS, Singer M. et al. Biomarkers for sepsis: more than just fever and leukocytosis—a narrative review. *Critical Care BioMed Central* 2022;**26**:1–31. <https://doi.org/10.1186/s13054-021-03862-5>.
 14. Chaudhry H, Zhou J, Zhong Y. et al. Role of cytokines as a double-edged sword in sepsis. *In Vivo* 2015;**27**:669–84.
 15. Manual U. Olink® NPX Explore HT. Olink proteomics AB. 2024. <https://olink.com/>.
 16. Van Eyk JE, Snyder MP. Precision medicine: role of proteomics in changing clinical management and care. *J Proteome Res* 2019[cited 2024 Aug 13];**18**:1. Available from: pmc/articles/PMC10372929/.