8-1-2021

Diagnostics and correction of batch effects in large-scale proteomic studies: a tutorial.

Jelena Čuklina

Chloe H Lee

Evan G Williams

Tatjana Sajic

Ben C Collins

See next page for additional authors

Authors

Jelena Čuklina, Chloe H Lee, Evan G Williams, Tatjana Sajic, Ben C Collins, María Rodríguez Martínez, Varun S Sharma, Fabian Wendt, Sandra Goetze, Gregory R Keele, Bernd Wollscheid, Ruedi Aebersold, and Patrick G A Pedrioli

Diagnostics and correction of batch effects in large-scale proteomic studies: a tutorial

Jelena Čuklina[1,2,3] , Chloe H Lee[1] , Evan G Williams[1,4] , Tatjana Sajic[1] , Ben C Collins[1,5] ,
María Rodríguez Martínez[3] , Varun S Sharma[1] , Fabian Wendt[6] , Sandra Goetze[6,7,8] ,
Gregory R Keele[9] , Bernd Wollscheid[6,7,8] , Ruedi Aebersold[1,10,*] & Patrick G A Pedrioli[1,6,7,8,**]

Abstract

Advancements in mass spectrometry-based proteomics have enabled experiments encompassing hundreds of samples. While these large sample sets deliver much-needed statistical power, handling them introduces technical variability known as batch effects. Here, we present a step-by-step protocol for the assessment, normalization, and batch correction of proteomic data. We review established methodologies from related fields and describe solutions specific to proteomic challenges, such as ion intensity drift and missing values in quantitative feature matrices. Finally, we compile a set of techniques that enable control of batch effect adjustment quality. We provide an R package, "proBatch", containing functions required for each step of the protocol. We demonstrate the utility of this methodology on five proteomic datasets each encompassing hundreds of samples and consisting of multiple experimental designs. In conclusion, we provide guidelines and tools to make the extraction of true biological signal from large proteomic studies more robust and transparent, ultimately facilitating reliable and reproducible research in clinical proteomics and systems biology.

Introduction

Recent advances in mass spectrometry (MS)-based proteomic approaches have significantly increased sample throughput and quantitative reproducibility. As a consequence, large-scale studies consisting of hundreds of samples are becoming increasingly common (Zhang *et al*, 2014, 2016; Liu *et al*, 2015; Mertins *et al*, 2016; Okada *et al*, 2016; Williams *et al*, 2016; Collins *et al*, 2017; Sajic *et al*, 2018). These technological and methodological advances, combined with proteins being the main regulators of the majority of biological processes, make MS-based proteomics a key methodology for studying physiological processes and diseases (Schubert *et al*, 2017). MS-derived quantitative measurements on thousands of proteins can, however, be affected by differences in sample preparation and data acquisition conditions such as different technicians, reagent batches, or changes in instrumentation. This phenomenon, known as "batch effects", introduces noise that reduces the statistical power to detect the true biological signal. In the most severe cases, the biological signal ends up correlating with technical variables, leading to concerns about the validity of the biological conclusions (Petricoin *et al*, 2002; Hu *et al*, 2005; Akey *et al*, 2007; Leek *et al*, 2010).

Batch effects have been extensively discussed, both in the genomic community that made major contributions to the problem about a decade ago (Leek *et al*, 2010; Luo *et al*, 2010; Chen *et al*, 2011; Dillies *et al*, 2013; Lazar *et al*, 2013; Chawade *et al*, 2014) and in the proteomic community which has faced the issue quite recently (Gregori *et al*, 2012; Karpievitch *et al*, 2012; Chawade *et al*, 2014; Välikangas *et al*, 2018). Nevertheless, finding solutions to the problem of batch effects is still a topic of active research. Although extensive reviews have been written on the topic (Leek *et al*, 2010; Lazar *et al*, 2013), researchers still get confused about the

1   Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland
2   PhD Program in Systems Biology, University of Zurich and ETH Zurich, Zurich, Switzerland
3   IBM Research Europe, Rüschlikon, Switzerland
4   Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Luxembourg, Luxembourg
5   Queen's University Belfast, Belfast, UK
6   Department of Health Sciences and Technology, Institute of Translational Medicine, ETH Zurich, Zurich, Switzerland
7   ETH Zürich, PHRT-CPAC, Zürich, Switzerland
8   SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland
9   The Jackson Laboratory, Bar Harbor, ME, USA
10 Faculty of Science, University of Zurich, Zurich, Switzerland
    *Corresponding author. Tel: +41 44 633 31 70; E-mail: aebersold@imsb.biol.ethz.ch
    **Corresponding author. Tel: +41 44 633 21 95; E-mail: pedrioli@imsb.biol.ethz.ch

terminology. For example, the distinction between normalization, batch effect correction, and batch effect adjustments is not always clear and these terms are often used interchangeably. To clarify how we use these terms in this Review, we compiled a glossary, found in Table 1. Some definitions are adapted from Leek *et al*, 2010.

There is also considerable debate on which batch correction method performs best, and multiple articles have compared various methods (Luo *et al*, 2010; Chen *et al*, 2011; Chawade *et al*, 2014). Other publications advise checking the assumptions about the data before selecting the bias adjustment method (Goh *et al*, 2017; Evans *et al*, 2018).

The issue of batch correction is further complicated by the fact that each technology faces different issues. Specifically, RNA-seq batch effect adjustment requires approaches that address sequencing-specific problems (Dillies *et al*, 2013). Similarly, MS methods in proteomics (e.g., data-dependent acquisition—DDA, data-independent acquisition—DIA, and tandem mass tag—TMT) also present several field-specific challenges. First, there is the problem of peptide to protein inference (Clough *et al*, 2012; Choi *et al*, 2014; Rosenberger *et al*, 2014; Teo *et al*, 2015; Muntel *et al*, 2019). As protein quantities are inferred from the quantities of measured peptides or even fragment ions, one needs to decide at which level to correct the data. Second, it is known that missing values can be associated with technical factors (Karpievitch *et al*, 2012; Matafora *et al*, 2017). Finally, when dealing with experiments with large sample numbers, typically in the order of hundreds, one needs to account for MS signal drift.

Here, we discuss the application of established approaches for batch effect adjustment. We also look at the methods that address MS-specific challenges. We start by providing an overview of the workflow and a definition of key terms for each step. In addition to

considering batch effect assessment and adjustment, we summarize the best practices for assessing the improvements in data quality post-correction. We also devote a section to the implications of missing values in relation to batch effects and potential pitfalls related to their imputation. We finish with a discussion and a future perspective of the presented approaches.

To facilitate the application to practical use cases, we illustrate all the relevant steps using three large-scale DIA and two DDA studies. For these "case studies", we primarily rely on the largest of the five datasets (i.e., Aging mouse study; preprint: Williams *et al*, 2021) and refer to the others where appropriate. The data analyses we show are only for illustration purposes and are not intended for deriving new biological insights.

Workflow overview

The purpose of this article is to guide researchers working with large-scale proteomic datasets toward minimizing bias and maximizing the robustness and reproducibility of results generated from such data. The workflow starts from a matrix of quantified features (e.g., transitions, peptides, or proteins) across multiple samples, here referred to as "raw data matrix" and finishes with "batch-adjusted" data, which are ready for downstream analyses (e.g., differential expression or network inference). We split the workflow into five steps, shown in Fig 1, and describe each of the steps below.

In the context of this article, we will use the term "adjust for batch effects" when referring to the whole workflow and "correct for batch effects" when referring to the correction of normalized data (see Table 1).

We provide a checklist that summarizes the most important points of the protocol in Table 2. It is also important to stress that batch factors should be already considered in the experimental design phase, to ensure that the data are not biased beyond repair, something that can happen when biological groups are completely confounded with sample preparation batches (Hu *et al*, 2005; Gilad & Mizrahi-Man, 2015). For an extensive discussion on experimental design, we refer the reader to previously published materials on the topic (Oberg & Vitek, 2009; Čuklina *et al*, 2020). Here, we assume that the experiment has been designed with appropriate randomization and blocking, ensuring the correctability of bias caused by batch effects.

In the accompanying "proBatch" package, we implemented several methods with proven utility in batch effect analysis and adjustment. We also provide tips for integrating other tools that might be useful in this context, and for making them compatible. proBatch is made available as a Bioconductor package (https://www.biocon ductor.org/packages/release/bioc/html/proBatch.html) and a pre-built Docker container (https://hub.docker.com/r/digitalproteomes/ probatch), as well as a GitHub repository (https://github.com/symb ioticMe/batch_effects_workflow_code) of the workflow with all code and data required to reproduce the case study analyses.

Extensive comparison of various methods has been published previously (Luo *et al*, 2010; Chawade *et al*, 2014), and here, we summarize the best practices from these papers, as well as reviews (Leek *et al*, 2010; Lazar *et al*, 2013) and application papers (Collins *et al*, 2017; Sajic *et al*, 2018), and turn them into principles that can guide the reader in choosing an appropriate methodology.

Table 1. Terminology.

Term	Definition
Batch effects	Systematic differences between the measurements due to technical factors, such as sample or reagent batches.
Normalization	Sample-wide adjustment of the data with the intention to bring the distribution of measured quantities into alignment. Most prominently, sample means and medians are aligned after normalization.
Batch effect correction	Data transformation procedure that corrects quantities of specific features (genes, peptides, metabolites) across samples, to reduce differences that are associated with technical factors, recorded in the experimental protocol (i.e., sample preparation or measurement batches). Usually samples are assumed to be normalized prior to batch effect correction. This step is often called "batch effect removal" or "batch effect adjustment" in the literature. Note the difference in the definition used here.
Batch effect adjustment	Data transformation procedure that adjusts for differences between samples due to technical factors that altered the data (sample-wise and/or feature-wise). The fundamental objective of the batch effect adjustment is to make all samples comparable for a meaningful biological analysis. In our definition, batch effect adjustment is a two-step transformation: first normalization, then batch effect correction. Performing normalization first helps feature-level batch effect correction by first alleviating sample level discrepancies.
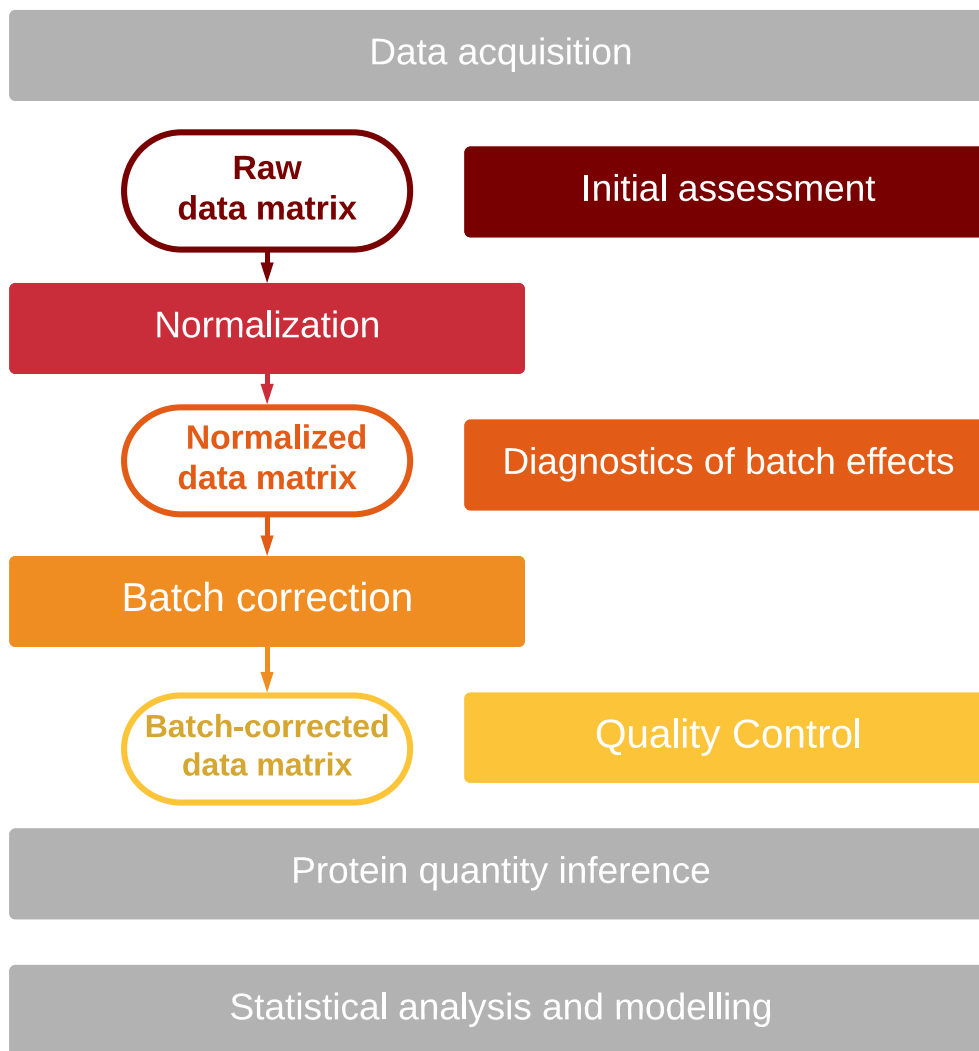
Figure 1. Batch effect processing workflow.

1. Initial assessment evaluates whether batch effects are present in raw data. 2. Normalization brings all samples from the dataset to a common scale. 3. Diagnostics of batch effects in normalized data. This step determines whether further correction is required. 4. Batch effect correction addresses feature-specific biases. 5. Quality control tests whether bias has been reduced while retaining meaningful signals.

Raw data matrix: choosing between protein/peptide/fragment level

This workflow starts with a raw data matrix, for which initial steps such as peptide-spectrum matching, quantification, and FDR control have been completed. Data are assumed to be log-transformed unless the variance stabilizing transformation (Durbin *et al*, 2002) is used. In the latter case, the data transformation is included in the normalization procedure.

We suggest performing batch effect adjustment on the peptide or fragment ion level, as this procedure alters feature abundances that are critical for protein quantity inference (Clough *et al*, 2012; Teo *et al*, 2015).

We also suggest that all detected peptides, including non-proteotypic peptides and peptides with missed cleavages, should be kept into consideration during batch effect adjustment. Keeping all measurements is required to better evaluate the intensity distribution within each sample, which is critical for subsequent normalization and correction steps.

Initial assessment

The goals of the initial assessment phase are to determine bias magnitude and sources and to select a normalization method. In most cases, the intensity distributions differ among samples. Comparing global quantitative properties such as sample medians or standard deviations helps with the choice of normalization methods and the identification of technical factors requiring further control.

Three approaches are particularly useful for initial assessment: (i) plotting the sample intensity average or median in order of MS measurement or technical batch, allows to estimate MS drift or discrete bias in each batch; (ii) boxplots allow to assess sample variance and outliers; and (iii) inter- vs. intrabatch sample correlation.

Table 2. Batch effect processing checklist.

Step	Substeps
Experimental design[a]	Randomize samples in a balanced manner to prevent confounding of biological factors with batches (technical factors).
	Consider adding replicates if possible, for example: (a) add replication for each technical factor; (b) regularly inject a sample mix every few (e.g., 10–15, but the exact number will need to be adjusted depending on experimental conditions) samples for control; (c) incorporate a sample mix per batch.
	Record all technical factors, both plannable and occurring unexpectedly.
Initial assessment	Check whether the sample intensity distributions are consistent.
	Check the correlation of all sample pairs.
	If intensities or sample correlations differ, check whether the intensities show batch-specific biases.
Normalization	Choose a normalization procedure, appropriate for biological background and data properties.
Diagnostics	Using diagnostic tools, determine whether batch effects persist in the data.
	Use quality control already at this step and skip the correction if it is not necessary.
	Tip: If the goal is to determine differentially expressed proteins, and the batch effects are discrete or linear, multi-factor ANOVA on normalized data is a sound statistical approach. This will adjust for batch effects while simultaneously identifying differentially expressed proteins. Note, that "hits" or differentially expressed proteins identified with this approach are valid even if diagnostic tools indicate the presence of batch effects. For more details on ANOVA methods, refer to (Rice, 2006).
Batch effect correction	Choose batch effect correction procedure, appropriate for the biological background and data properties, especially those detected at the previous step.
	Repeat the diagnostic step.
	Assess the ultimate benefit with quality control.
Quality control	Compare correlation of samples within and between the batches. Pay special attention to replicate correlation, if these are available.
	Compare correlation of peptides within and between the proteins.

[a]For details on experimental design, see (Čuklina et al, 2020).

A higher correlation of samples from the same batch compared with unrelated batches is a clear sign of bias. Optionally, a few proteins or peptides can be checked for signs of bias.

Normalization

The goal of normalization is to bring all samples to the same scale to make them comparable. Commonly used methods of normalization are quantile normalization, median normalization, and z-transformation. Two main considerations drive the choice of normalization method:

- Heterogeneity of the data: If samples are fairly similar, the bulk of the proteome does not change, and thus, techniques such as quantile normalization (Bolstad et al, 2003) can be used. In datasets in which the samples are substantially different (i.e., when a large fraction of the variables are either positively or negatively affected by the treatment) different methods, such as HMM-assisted normalization can be used (Landfors et al, 2011). Additionally, if some samples are expected to have informative outliers (e.g., muscle tissue, in which a handful of proteins are several orders of magnitude more abundant than the rest of the proteome), methods that keep the relationship of outliers to the bulk proteome need to be used (Wang et al, 2021).
- Distribution of sample intensities: The initial assessment step, especially boxplots, indicates which level of correction is required: In most cases, shifting the means or medians is enough, but when variances differ substantially, these need to be brought to the same scale as well.

It should be noted that after normalization, no further data correction might be required. This can be determined with the diagnostic plots and quality control methods described below. If the results are satisfactory, keeping data manipulation minimal is advisable.

Diagnostics of normalized data

While normalization makes the samples more comparable, it only aligns their global patterns. Therefore, batch effects affecting specific proteins or protein groups might still represent a major source of variance even after normalization. Thus, the diagnosis of batch effects is most informative when performed on normalized data.

The diagnostic approaches can be divided into proteome-wide and peptide-level approaches. The main approaches for proteome-wide diagnostics are as follows:

- Hierarchical clustering is an algorithm that groups similar samples into a tree-like structure called a dendrogram. Similar samples cluster together, and the driving cause of this similarity can be visualized by coloring the dendrogram by technical and biological factors. Hierarchical clustering is often combined with a heatmap, mapping quantitative values in the data matrix to colors which facilitates the assessment of patterns in the dataset.
- Principal Component Analysis (PCA) is a technique that identifies the leading directions of variation, known as principal components. The projection of data on two-component axes visualizes sample proximity. Additional coloring of the samples by technical/biological factors, or by highlighting replicates, facilitates the interpretation of what drives sample proximity. This technique is particularly convenient to assess clustering by biological and technical factors or to check for replicate similarity. Visualization without sample point or label overlay effects works in our experience up to about 50–100 samples in a dataset.

One should be careful in interpreting proteome-wide diagnostics because these methods were designed for data matrices without missing values. Proteomic datasets often contain missing values for

Box 1. Missing values

Proteomic experiments now routinely profile hundreds or thousands of proteins across hundreds of samples. However, detecting all proteins without missing values across the whole dataset is not yet feasible. The patterns of "missingness" are known to be batch-specific (Karpievitch *et al*, 2012), and some workflows are susceptible to a rapid inflation of missing values as the number of batches increases (Brenes *et al*, 2019). This is also true for the largest datasets of this manuscript: aging mouse DIA and TMT datasets (see Box 1 Figure, Figs EV5 and EV6 for details).

It should be noted, that even though "missingness" for low-abundant peptides is more common (i.e., an issue related to the dynamic range and sensitivity of the mass spectrometer), this problem can also arise due to fundamental peptide interference regardless of their abundance or the acquisition parameters.

Missing values can also affect batch effect correction methodologies. For instance, the current implementation of ComBat (Johnson *et al*, 2007) does not work if a peptide is missing in one batch. One possible solution is to remove all peptides with missing values before the batch correction (Lee *et al*, 2019). However, this may lead to loss of valuable quantitative information. Thus, methods which are more robust to missing data, such as median centering, can sometimes be better suited for proteomic data.

Missing values are often imputed, by filling them with zeros, random small values (Tyanova *et al*, 2016) or re-quantification of elution traces (Röst *et al*, 2016). Such imputation, however, can introduce bias that is batch- or peptide-specific, as seen in Figs EV5A and EV6. In turn, this skews batch effect diagnostic methods, such as hierarchical clustering, PCA, or PVCA. In these cases, batch effect assessment will be biased, as the clustering pattern will be driven by missing values (Fig EV5A). One can estimate this effect by varying the fraction of missing values and assessing to what extent the batch effects are driven by consistently quantified peptides vs. missing values containing ones (Fig EV5B).

More importantly, imputed values bias the analysis past the batch effect adjustment stage. As shown in Box 1 Figure B and C, if re-quantifications ("with requants") values inferred from MS elution traces are used, the correlation within batches seems higher than the correlation of replicates, while this problem is not observed when imputation is not used ("no requants"). Protein inference is also affected by the imputation on lower levels.

Finally, provided that there are enough confidently quantified values, many downstream analysis techniques, such as differential expression or protein correlation analyses, can handle missing values. We therefore advise to avoid imputation, or at least suggest to perform it after batch correction whenever possible.
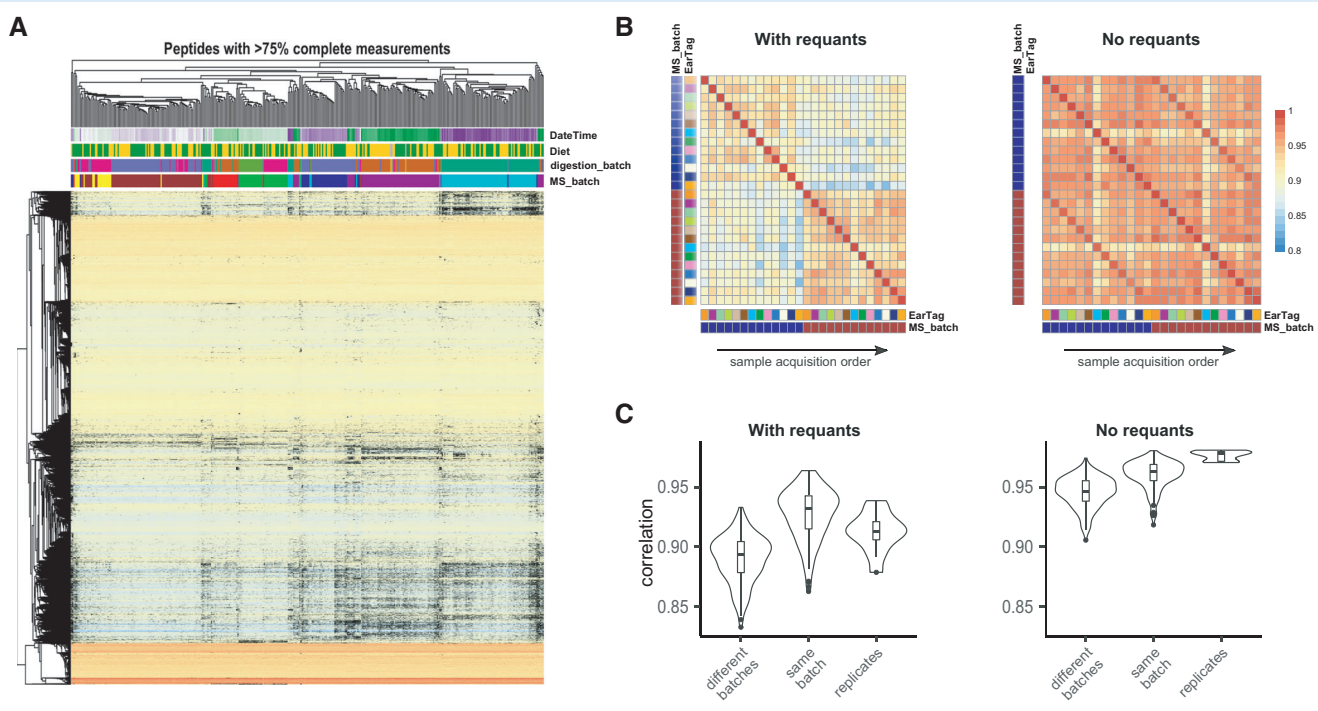
Box Figure 1. The problem of missing values in batch effect diagnosis and correction: Aging mouse study. (A) Hierarchical clustering and heatmap of normalized data; missing values shown in black. The missing values are non-randomly associated with the batch; (B) heatmap of selected sample correlation: Stronger correlation of samples within Batch 2 (blue) and Batch 3 (brown) is visible in the data with "requants", and replicate correlation is much more prominent in the data without "requants"; (C) distribution of selected sample correlation: same effect, as in (B) showing the distribution of sample correlation.

technical or biological reasons. For more details, we refer the reader to Box 1.

In proteomics, peptide-level diagnostics are as useful as proteome-wide diagnostics. As in other high-throughput measurements, individual features, in this case, peptides, are visualized to check for batch-related bias. In proteomic datasets, spike-in proteins or peptides can be added as controls. In most DIA datasets, iRT peptides (Escher *et al*, 2012), if added in precise concentrations, are well suited for individual feature diagnostics. It should be noted that individual peptides have a variety of different responses to various

batch effects, so checking a handful of peptides is necessary, whether endogenous or spiked-in.

Another reason to check individual peptides in proteomics is to examine the trends associated with sample running order. These trends might occur as MS signal deteriorates and require special correction approaches.

Note, that in proteomics, individual features are sometimes not peptides, but transitions or peptide groups. Thus, methods referred here as peptide-level diagnostics are applicable to any feature-level diagnostics.

Batch effect correction

Diagnostics help to determine whether batch effect corrections are needed. While global sample patterns are corrected during normalization, batch effects affect specific features and feature groups, and that is the level on which they need to be corrected.

In proteomic datasets, two types of batch effects are frequently encountered, continuous and discrete. If batch effects are continuous, e.g., manifest as MS signal drift progressing from run to run during the sample measurement process, an order-specific curve needs to be fitted, such as a LOESS fit, or by using any other continuous algorithm. Signal drifts are likely to occur in studies profiling hundreds of samples. This problem is more prominent in mass spectrometry as compared to next-generation sequencing and is thus still relatively new to the research community.

Discrete batch effects manifest as feature-specific shifts of each batch as a whole. Here, methods such as mean and median centering work very well. An advanced modification of the mean shift is provided by ComBat (Johnson *et al*, 2007) that uses a Bayesian framework which can be applied to proteomic data (Lee *et al*, 2019). However, ComBat requires that all features are represented in each of the batches. Therefore, especially in large-scale proteomic datasets, applying ComBat might require the removal of a substantial number of peptides that happen to be missing in at least one batch, regardless of how small this batch is (see Box 1 for details). Thus, one should be very careful when choosing the method for batch effect correction.

Quality control

The purpose of the quality control step is to determine whether the adjustment procedures—normalization and/or batch effect correction—have improved the data. At this step, the data after adjustment are compared with the raw data matrix. There are two types of criteria to evaluate the data quality: (i) removal of the bias (negative control) and (ii) improvement of the data (positive control).

Typically, bias is considered removed if the similarity between samples is no longer driven by technical factors. This means that neither hierarchical clustering nor PCA shows clustering by batch, and the correlation of samples from the same batch is no longer stronger than the correlation of unrelated samples. Also, individual features should not show batch-related biases. Thus, comparison of diagnostic plots for raw and adjusted data serves as the negative control.

Proving improvement achieved by batch correction is much harder. It is common to take "improved clustering by biological condition" or "higher number of differentially expressed proteins" as a positive control and generally, as a sign of data quality enhancement. However, both criteria are subjective: It is impossible

to know beforehand, whether biological groups are separable in the proteomic space, especially if only a subset of proteins changes while the bulk of the proteome does not. Similarly, it is not possible to predict whether higher sensitivity for differential expression comes at the expense of added false-positive hits. Therefore, we do not recommend using these criteria to assess normalization or batch effect correction. As described above, the choice of the method should rather be based on the properties of the samples. In general, since batch adjustment removes a certain portion of variance, the coefficient of variation for peptides and proteins in replicated samples should decrease. This is especially true for spike-in peptides or proteins that are added to samples in controlled quantities. A stronger positive control is the assessment of reproducibility, such as comparison of lists of differentially expressed proteins or regression/classification models derived from two or more sample sets belonging to different batches (Lazar *et al*, 2013). It is expected that in adjusted datasets, the resulting lists of differentially expressed proteins, or proteins providing optimal class separation, will be highly overlapping (Shabalin *et al*, 2008). If two sample sets are independently used for predictive modeling, the predictive performance of such models is also expected to be comparable in adjusted datasets (Luo *et al*, 2010). Note, however, that while this method generalizes well to studies with data acquired by different technologies (e.g., microarrays vs. RNA-seq for transcriptomics or DIA vs. DDA for proteomics), it is restricted to fairly large datasets (several dozens, preferably, hundreds of samples), as predictions from small-scale experiments tend to be unstable (underpowered).

Here, we also propose two positive control methods that do not rely on large sample size and are applicable to most proteomic experiments. The first is based on sample correlation. It is expected that the correlation between technical or biological replicate samples is higher than the correlation of unrelated samples. Particularly, the distribution of replicate correlations should be clearly shifted upwards, even though replicates might occasionally correlate less than some unrelated sample pairs, and this distinction should be strengthened by batch adjustment procedures. Following similar logic, other distance metrics can be used to assess sample proximity, which can also be visualized as improved clustering of replicated samples, seen on hierarchical clustering or PCA component plots. The latter method visualizes every sample in the experiment and thus is most suited to assess studies with up to 150 samples, while bigger sample sizes are harder to visualize. The second assessment method is specific for bottom-up proteomics and makes use of peptide correlation. Correlation of unrelated peptides is expected to be close to zero, while peptides originating from the same protein are likely to be positively correlated. Since tens of thousands of peptides are routinely detected in modern high-throughput proteomic experiments, improvements in this metric are a reliable readout of data quality following batch adjustment.

Description of datasets used to illustrate the workflow

To illustrate the application of the workflow described above, we use five proteomic datasets (three acquired in DIA and two in DDA mode), described in Table 3.

The first study, called here "InterLab study", assessed the robustness of SWATH-MS in a multi-lab setting (Collins *et al*, 2017; Data

Table 3. Dataset description.

Sample	Organism	Sample source	Sample-to-sample heterogeneity	Technical factors	Biological factors	Protein (peak groups/ precursors) number	Number of samples	Dataset accession
InterLab study	Human	Cell culture	Very low: samples come from the same tissue cultures and differ only by few spike-in peptides	Data acquisition sites Profiling days	None	4,077 (31,886)	229	PRIDE PXD004886
PanCancer study	Human	Blood	High: samples come from cancer patients and matched controls with different cancer localization	Protein digestion batch	Case / control Cancer localization	205 (1,360)	162	PRIDE PXD004998
Aging mouse study	Mouse	Liver tissue	Medium: samples come from population of inbred mice originating from two parental strains	Protein digestion batch MS batch MS drift	Strain Diet Age	3,940 (32,449)[a]	413	PRIDE PXD009160
TMT mouse study	Mouse	Liver	Medium: samples from a population of inbred mice originating from eight parental strains	Sample preparation batch MS batch MS drift	Strain Age	6,813(66,418)[a]	120	PRIDE PXD018886
Bariatric surgery study	Rat	Lymph	Medium: samples come from inbred rat population	Liquid handling robot	Gastric bypass vs. placebo surgery	302 (1,987)	68	MassIVE MSV000087519

[a]Number of proteins and peptides before filtering for peptides with too many missing values.

ref: Collins *et al*, 2017). A set of 30 stable isotope labeled (SIL) peptides (Ebhardt *et al*, 2012), partitioned in five groups, was serially diluted in HEK293 cell lysate. The SIL peptides in the resulting samples spanned a concentration range from 12 amol to 10 pmol. These five sample sets were distributed to 11 laboratories worldwide for measurement by SWATH-MS according to a predetermined schedule. Each of the samples was run on 3 separate days, with the exception of the 4th sample that was run three times on each day. In total, 229 samples were profiled. Thus, the technical covariates whose effect needed to be assessed were the data acquisition site and day. Note that due to the technical nature of the study, no biological signal needed to be identified. As only a small number of SIL peptides is different across these samples, all changes can be attributed to technical covariates, and therefore, the samples in this study can be treated as replicates. Within this manuscript, we analyze only the influence of the acquisition site as a batch factor.

The second study, named here "PanCancer study" (Sajic *et al*, 2018; Data ref: Sajic *et al*, 2018), profiled the blood plasma glyco-proteome of a cohort of patients with five solid carcinomas and matched controls. In total, 155 blood plasma samples were collected. Protein digestion and glycopeptide enrichment were performed in 4 batches, several weeks apart. To account for sample preparation reproducibility, 7 biospecimens were replicated and allocated to a different batch. To control for intra-sample variation caused by the sample preparation protocol, bovine fetuin-B was spiked in equal amounts into each plasma sample. In total, 162 samples were measured (the validation cohort from the original manuscript is omitted from this analysis).

The third dataset is called here the "Aging mouse study" (preprint: Williams *et al*, 2021; Data ref: Williams *et al*, 2021). In this study, 413 liver proteomes were measured from 341 individual mice from the BXD reference mouse population (Peirce *et al*, 2004) to identify changes associated with age. Similarly to prior BXD mice metabolic profiling experiments (Williams *et al*, 2016), genetically identical cohorts of animals were also subjected to either chow or high-fat diet. The samples were randomized with respect to biological covariates (age, diet, sex), and samples from two mice with EarTags "ET1506" or "ET1524" were both injected 10 times at various intervals throughout the run to control for signal consistency. Additionally, a mix of samples was shot 3 times as control. In this experiment, two technical factors are known to affect the measurement. First, the samples were digested in five batches. Second, to compensate for signal deterioration, MS data acquisition was interrupted for machine cleaning and tuning, resulting in 7 mass spectrometry batches. These are shown as vertical lines in Fig 2. Except for replicates, samples were run in the same order of digestion batches (see Fig EV1A), so these two factors are mostly confounded (i.e., digestion and MS batch mostly overlap). Therefore, in our analysis we only correct for MS batch, unless noted otherwise. Given the particularly severe MS signal deterioration at the end of MS batch 2, the last 13 samples of this batch were profiled again as first samples of MS batch 3. In total, 375 proteome acquisitions were considered in this manuscript, while 38 acquisitions were discarded prior to analysis due to major acquisition failures. See also "Appendix" in supporting information for a summary of the original experimental setup of this study.
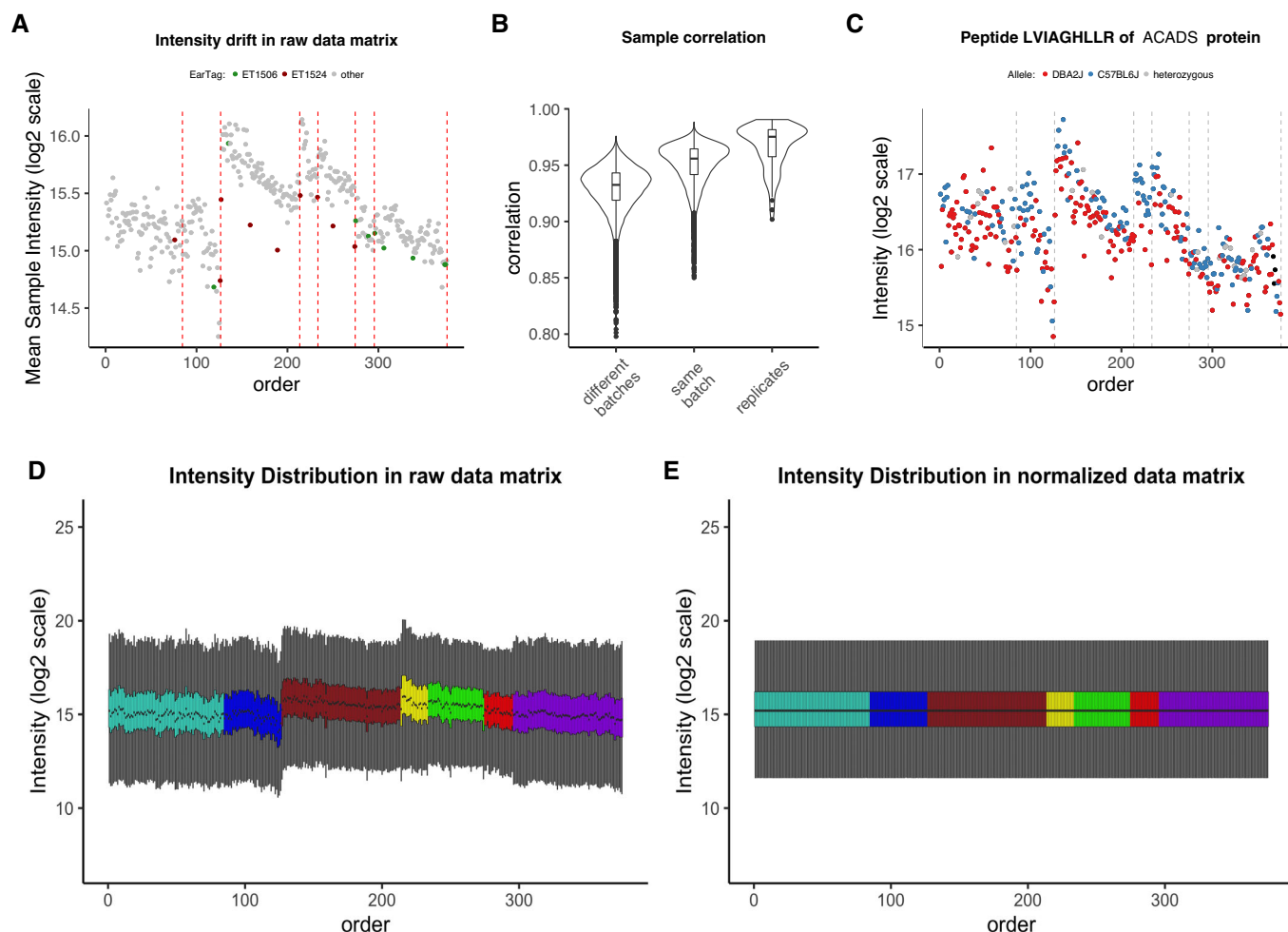
Figure 2. Initial assessment and normalization of the Aging mouse study.

(A) Mean intensity in raw peptide matrix vs. sample running order with repeatedly replicated samples shown in color. Vertical dotted lines indicate MS batch boundaries; (B) distribution of unadjusted sample intensity correlations—between batches, within batches, and in replicated samples; (C) bias in protein quantification: representative ACADS protein, the quantity of which follows the drift of the average sample intensity, preventing allele separation and QTL detection; (D) boxplots of sample intensities in raw, unnormalized peptide; (E) boxplots of sample intensities after quantile normalization. All plots represent peptide-level data.

The fourth study, named here "TMT mouse study", used data acquired from livers from 120 individuals from the Collaborative Cross reference mouse population, taken at 8 weeks of age (preprint: Keele *et al*, 2020; Data ref: Keele *et al*, 2020). Data were acquired in 12 TMT batches of 10 samples each, with a six-month gap between batches 10 and 11. Batches 1–10 and 11–12 were prepared and run directly sequentially. The peptide measurement table from that paper was used as input for proBatch, and *cis*-pepQTLs were calculated before and after proBatch.

The fifth study, named here "Bariatric surgery study" (Kaufman *et al*, 2019; Data ref: Kaufman *et al*, 2019), profiled N-linked glycopro-teomes from rat lymph before and after gastric bypass surgery. The cohort consisted of 68 lymph samples originating from rats before and after gastric bypass surgery (RYGB) or placebo surgery (SHAM). Samples were collected from rats before, 5, 10, and 21 days after the operation. The samples for this study were processed using a Versette automated liquid handling system (ThermoFisher Scientific) in a

96-well plate format. Differences in the performances of Versette's channels have introduced a "robot batch" into these data. The samples were measured in label-free DDA mode, and glycopeptides were quantified using Progenesis (non-linear dynamics).

All in all, these datasets are representative of various applications of large-scale proteomic studies. They make use of different sample sources (i.e., cell cultures, patients, model organisms). They ask technical and biological questions about proteomes of varying complexity and present different degrees of sample-to-sample heterogeneity. In this respect, the InterLab study is very homoge-neous, to the point that all the samples are essentially technical replicates. The PanCancer study is highly heterogeneous and comprises 205 proteins identified in samples originating from different hospitals and different tissues. Finally, the model-organism-based studies represent an intermediate case. On the one hand, its subjects were genetically related, but on the other hand, sampling introduces a certain amount of sample heterogeneity.

Case studies

Initial assessment

The main goal of the initial assessment is to set a baseline for the magnitude and nature of batch effects in a particular dataset. At this stage, the data matrix is "raw", and the quantities are reported as measured, without any calibration, normalization, or correction with regard to their values in other samples.

Thus, it is essential to get a quick overview of the data by comparing global statistics, such as the average intensity, or correlation of samples and of few individual proteins with known expected abundance. In mass spectrometry, it is helpful to first plot these statistics by sample running order, as it is common for the measured signal to drift (e.g., due to deterioration of the LC and/or the MS). This is clearly seen in Fig 2A, where the average intensities of samples from the Aging mouse study are plotted vs. the sample running order. In this case, the intensity tended to deteriorate after 50–70 samples. The resulting interruptions for cleaning and calibrating the instrument determined the discrete mass spectrometry batches. As this type of bias is predictable, but cannot be entirely planned for in advance, it is particularly important to randomize the samples and to include replicates (for more details on sample replication in this dataset, see the "Description of datasets used to illustrate the workflow" section).

Not only do batch effects introduce shifts in the total sample intensity distributions, but they also lead to a spurious correlation between features (i.e., fragments, peptides, or proteins). It is common, for samples belonging to the same batch to have strong correlation (as seen in Figs 2B, EV1A, and EV2B). Often this correlation is not only stronger than the correlation of samples from different batches ("between batches") but also stronger than the correlation of replicates, hence the importance of assessing correlation distributions as early as possible. Sample correlation can be visualized as a square heatmap (Aging mouse—Fig EV1A, InterLab study—Fig EV2B) or as a correlation distribution box/violinplot (Fig 2B). The former is preferred with smaller sample sizes (i.e., roughly < 150 samples) and the latter with large datasets, as it is not possible to assess replicate correlation in large datasets on the heatmap.

Optionally, one can complement the initial assessment with the analysis of a few specific features (peptides or proteins), for which prior information is known. Here, we plot a peptide from a representative protein *ACADS* with a known genetic sequence variant affecting its own expression at a quantitative trait locus (cis-QTL), which can be seen as the bimodal expression separation on the plot according to the two possible ACADS alleles of the inbred strain. However, as seen in Fig 2C, the intensity drift, similar to that in Fig 2A, prevented the corresponding alleles in this BXD population, which descends from two strains, DBA/2J (red) and C75BL/6J (blue), from being separated based on expression level.

Last but not least, intensity boxplots are extremely powerful at this stage of the analysis, as they visualize in a single plot: median, quantiles, and outliers. This allows one to see whether there are batch-specific intensity patterns such as shifts (as in the InterLab study, see Fig EV2A), or drifts (as in Fig 2D), or no evident batch-associated patterns (PanCancer data; Fig EV3A). To detect the patterns, the samples should be sorted by running order or by batch (if the batches, such as digestion batches have been randomized prior

to MS analysis). Note that in some cases, intensity patterns might be easier to spot on average intensity plot (compare to Fig 2A).

Normalization

Normalization is an essential step in removing bias from the data as it brings the samples to the same scale, making the measured quantities comparable (Leek *et al*, 2010). As stated in the "Workflow overview" section, the choice of normalization method should take two factors into account: (i) heterogeneity, as assessed from previous knowledge, and (ii) global quantitative sample properties, e.g., mean/median/variance, as indicated by the initial assessment.

Quantile normalization (Bolstad *et al*, 2003) is applicable to a wide range of samples and was chosen for the aging mouse data, the PanCancer dataset, and the Bariatric surgery study. As shown in Figure 2D and E (and Fig EV3A and B), the intensity distributions after quantile normalization are very similar. This is desirable in experiments where the majority of features are not expected to change but can be problematic in experimental setups where outliers bear important information (Wang *et al*, 2021). The TMT mouse data were normalized outside of proBatch (preprint: Keele *et al*, 2020). Importantly, proBatch allows data to be taken up at different steps of the batch effects processing workflow and was used for this dataset only post-normalization.

Median centering normalization only brings the medians to the same scale and thus is a "milder" approach. This normalization was chosen for the InterLab study and brought the peptide quantities measured at different sites much closer to each other (see Fig EV2C). In this dataset, shifting the medians to the same value was sufficient to remove a substantial portion of bias (see Fig EV2D, note that the correction on fragment level has residual median discrepancies on protein level). The resulting improvement in signal quality is also reflected in the improved protein quantification precision as illustrated by the reduction in coefficient of variation for the majority of proteins in the dataset (see Fig EV2E).

As shown in this example, normalization alone is sometimes the only required correction. In the median-normalized InterLab study, the improvement of quantification of spike-in peptides (Fig EV2C) and the decrease in protein coefficient of variation (Fig EV2E) serve as quality controls and indicate that further batch correction steps can be skipped.

This is of course not always the case, especially for larger datasets, and batch effect diagnostics should be used, as presented in the next section, to determine whether further correction steps are required.

Diagnostics

Normalization harmonizes overall sample intensities; however, batch effects can still be present at the level of specific features and bias the quantities of many peptides and proteins. Various methods can be used to assess the extent of remaining bias in the data. Most methods characterize the dataset as a whole, thus working on the proteome level. Methods such as Principal Components Analysis (PCA) and Hierarchical Clustering are the best established. For instance, in the Aging mouse study, the following patterns became apparent (see Fig 3A): First, we see a strong clustering of samples by mass spectrometry batch; second, while replicates tend to be close, they are not necessarily the samples with the smallest distance to each other. An overlaid visualization of other factors,
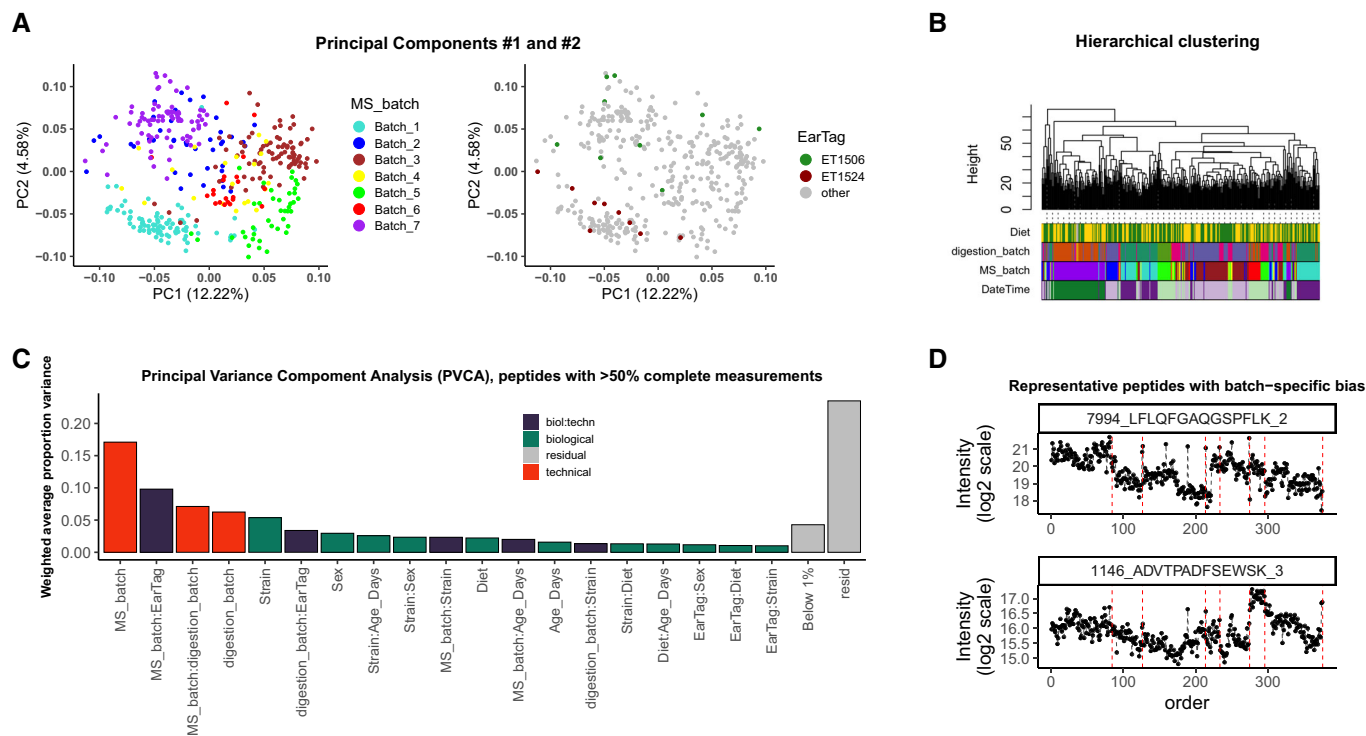
Figure 3. Diagnostics of batch effects: Aging mouse study.

(A) Principal Components #1 and #2 colored by MS batch (left) and replicates (right), percentage of variance in each PC shown in brackets. The effect of clustering by MS batch is dominating, but the replicated samples are closer to each other than just random samples; (B) hierarchical clustering of samples, with leaves colored by diet, digestion batch, MS batch, and date–time of sample acquisition is dominated by technical factors; (C) Principal Variance Component Analysis of peptides, detected in >50% of samples demonstrates, that the technical factors, such as MS batch, digestion batch, and their combination, have a profound effect on the data, while biological factors such as strain, sex, and age account for a much smaller fraction of variance; (D) peptide-level plots for two iRT peptides demonstrate that batch effect manifests also as MS signal drift that requires correction.

such as digestion batch, diet, or acquisition date, can be found in Fig EV1B. Also for the TMT mouse study, PCA shows strong clustering by batch (TMT batch, see Fig 4A).

Hierarchical clustering, in contrast, allows one to visualize multiple factors at once (Fig 3B). Overall, the same patterns identified in the PCA plots are confirmed as follows: Clustering is driven primarily by MS and digestion batch, while the effect of diet, a dominant biological factor, is not as strong. Similarly, in the PanCancer study (Fig EV3C and D, top parts), we see that the digestion batch is the key clustering factor, and also in TMT mouse study, TMT batch (Batch) drives the clustering of samples (Fig 4B).

Principal Variance Components Analysis (PVCA) transforms the intuition of visualization brought about by PCA and Hierarchical Clustering into numbers. The weights of each factor in each Principal Component are combined, thus providing a concise summary of variance distribution between biological and technical factors and their combination. In the aging mouse dataset, we see in Fig 3C that technical factors such as MS batch, digestion batch, and their combination are leading drivers of variance. Whereas biological factors, such as diet, sex, and age, are much less prominent. It should also be noted that a substantial fraction of variance is "residual" and cannot be explained by the annotated sample characteristics. PVCA also shows a clear batch effect from the liquid handling robot for the Bariatric surgery study (see Fig EV4A).

It is important to point out that PCA, hierarchical clustering, and PVCA require matrices with complete feature measurements, or with imputed missing values. Since in MS proteomic missing values are often batch-specific, this might lead to an overestimate of the batch-related clustering (variance explained) and to over-pessimistic assessment of data quality (see also Box 1).

In addition to proteome-wide diagnostics, it is often informative to visualize a few individual features to assess for the presence of feature-level bias. Spike-in peptides and proteins are particularly handy for this type of diagnostics. We illustrate this in the aging mouse data using iRT peptides that were spiked at constant quantities in all samples. In Fig 3D, one can see, that despite the normalization, batch-specific bias still affects the signal. Moreover, this bias is order-related and manifests differently for each of the peptides (e.g., in samples 127–212 of MS_batch 3). In the PanCancer study, we see that protein quantities are biased by batch effect on the example of the spiked-in bovine protein fetuin-B (Fig EV3E and F). In contrast to the Aging mouse study, here there is no order-related drift, but rather each batch mean is shifted. In conclusion, feature-level bias can be very different in each dataset and checking several representative peptides can help understand the exact nature of the bias.

Together, proteome-level and feature-level diagnostics guide the choice of appropriate batch correction procedure.
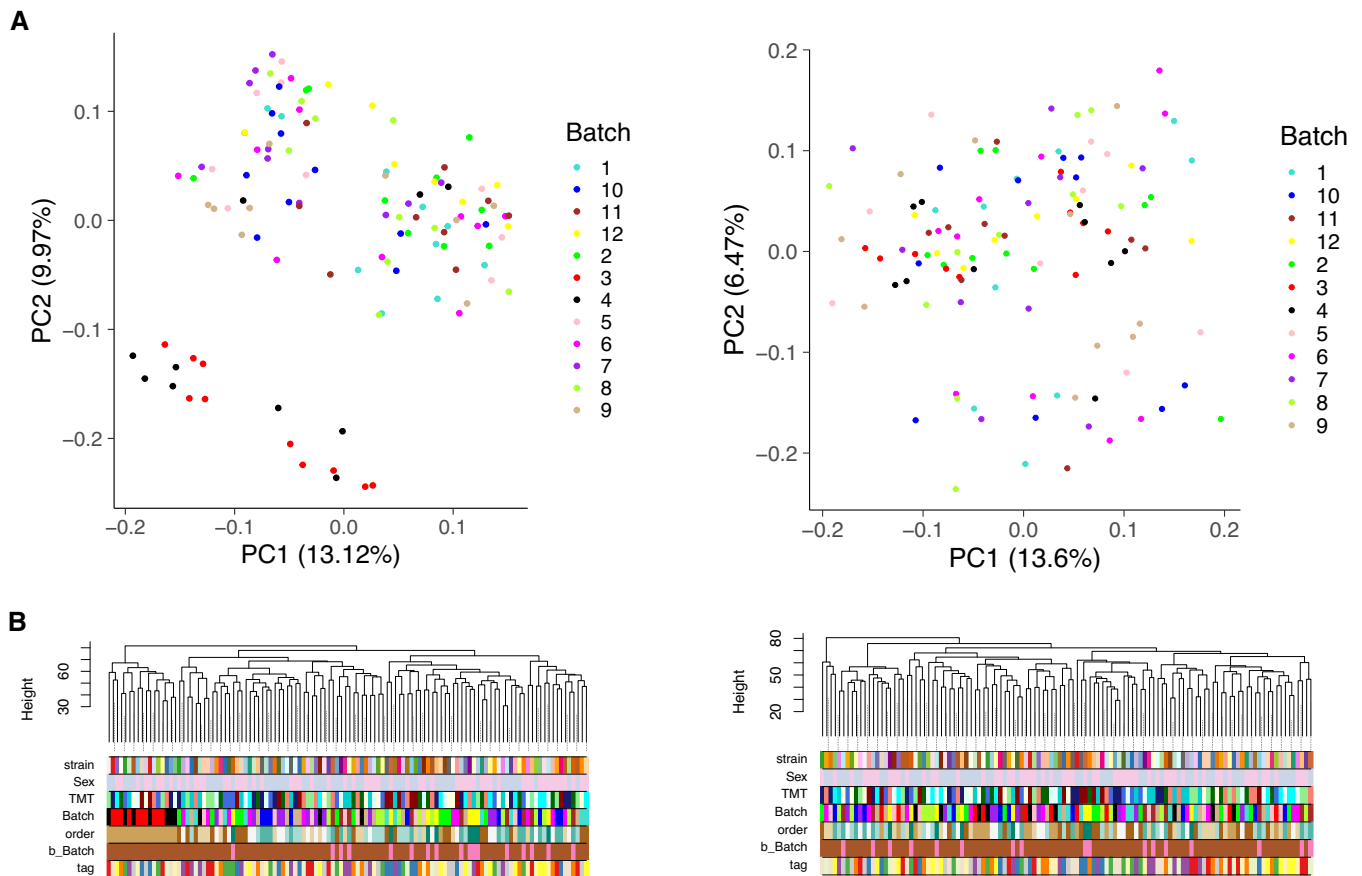
Figure 4. Batch effects in the TMT mouse study.

(A) PCA and (B) Hierarchical clustering in matrix of peptides with complete measurements before (left) and after (right) correction.

Batch effect correction

The goal of batch effect correction is to alleviate the residual bias after normalization. In most cases, normalization significantly reduces the unwanted variance (see representative peptides in Fig EV1C) and, as in the InterLab study, might sometimes be the only required adjustment. However, in many cases, the batch effect diagnostic procedures described in the previous section will reveal that individual features are differentially affected by batch effects even after normalization (Figs 3D and EV3E). Batch effect correction is typically applied at the feature level.

Feature-level bias in proteomics can have different roots. In the aging mouse data, the peptides manifest an order-specific, continuous bias, while data in the PanCancer exhibit discrete shifts. The latter can be corrected with methods established by the genomic community, whereas the former is typical of large MS-based proteomic experiments and will therefore be discussed in more detail in the following paragraphs.

In MS-based proteomics, samples are analyzed sequentially, typically using an online chromatographic system directly coupled to a mass spectrometer. Various components of this system are susceptible to degradation of performances over time (e.g., changes in the properties of the chromatographic material, emitter degradation, contamination of MS lenses, mass calibration drift). Various

factors, such as sample quality and composition, can contribute to the speed of this degradation. This is particularly relevant in large-scale studies.

To address this complex bias, we have developed a two-step batch correction procedure shown for the aging mouse in Fig 5. In the first step, MS signal drift is corrected based on non-linear curve fitting. In the second step, remaining discrete batch effects are addressed using a discrete batch correction procedure (e.g., median centering, ComBat).

As a curve-fitting algorithm for the first step, we have chosen LOESS, as it combines computational simplicity with relative flexibility of the fit characteristics. The procedure runs as follows: For each peptide and each batch, a unique curve is fit (Fig 5A) and then subtracted from the normalized intensity value in each sample, leading to measurements, whose median is different in each batch (Fig 5B). Note, that LOESS has a parameter "span": Smaller span fits minute details, while when the span is too big, some of the trends get missed, meaning each time the span has to be adjusted to each dataset individually (see Fig EV1D). Alternatively, other algorithms, such as SVM and Random Forest, can be used to correct for the MS signal drift (Shen *et al*, 2016; preprint: Luan *et al*, 2018). For the second step, the residual discrete batch effects were removed by median centering (Fig 5C).
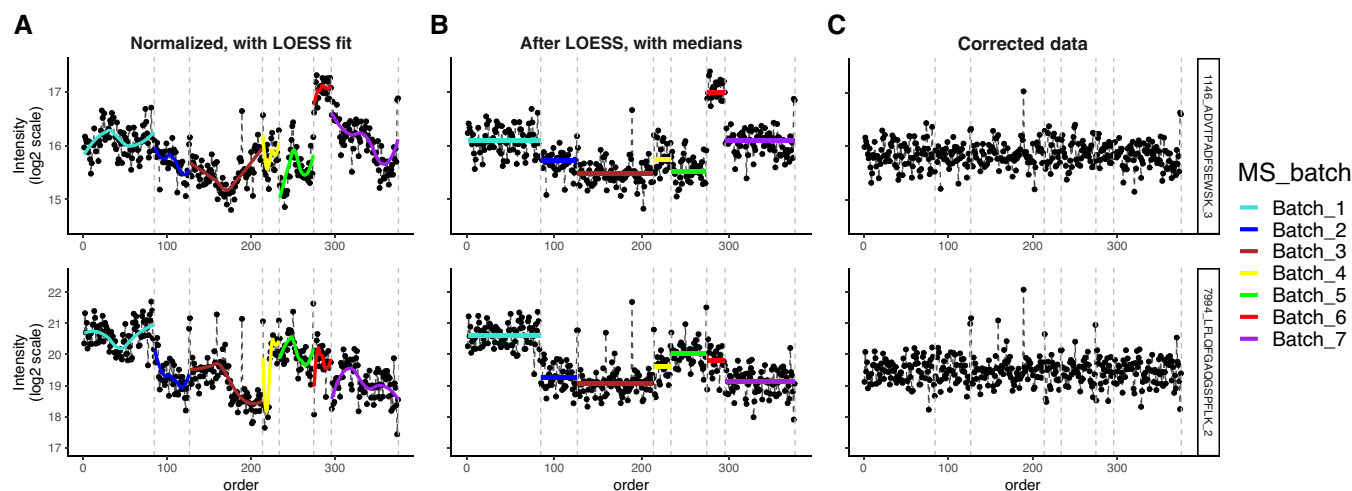
Figure 5. Two-step correction of batch effects.

(A) Fitting a LOESS curve for every peptide in each batch and subtracting the fit; (B) using medians for correction of the residual discrete batch effect; (C) the corrected data are more uniform and can now be used for downstream analysis. Data from the Aging mouse study.

Datasets that do not exhibit MS signal drift require only discrete batch correction—second step of the aging mouse dataset. Here, the PanCancer (Fig EV3C–F) and TMT mouse (Fig 4A and B) were corrected in a single step using median centering. Similarly, in the Bariatric surgery dataset (Fig EV4A and B) the liquid handling robot represented the main source of discrete bias and was corrected in a single step using ComBat (Johnson et al, 2007).

When several batch factors affect the data (as diagnosed by PCA, hierarchical clustering, and/or PVCA), all factors should be accounted for during correction. This typically means that batch factors get combined. This, however, is not always possible (e.g., batches become too small). In the Aging mouse study, for instance, the two main technical factors, MS batch and digestion batch, were highly confounded. Thus, we opted to only correct for the MS batch.

In conclusion, batch effect correction removes variance from known, annotated batch factors. Whether the adjustment procedures —normalization and batch effect correction—improved the data signal, remains, however, to be determined by a quality control step.

Quality control

The main goal of the quality control step is to determine whether the data quality has improved.

Examples of negative controls are the plots discussed in the diagnostic sections such as individual peptide plots, PCA, hierarchical clustering, or PVCA. In most cases, these show that samples are no longer affected by batch-specific patterns (e.g., PanCancer study hierarchical clustering in Fig EV3C and D and spike-in protein in Fig EV3E and F).

Depending on the experimental setup, one can sometimes judge the data improvement through the presence of a better biological signal. In the Aging mouse study, this can be achieved via mapping protein quantitative trait loci (pQTL), particularly cis-QTLs (i.e., alleles of specific genes have DNA variants which affect their own protein's expression level, which is tested on peptide level). As seen in Fig 6A, certain proteins become clearly easier to separate

according to their variant allele and less biased than before the bias adjustment (compare to Fig 2C). As a result, there is a clear increase in pQTL detection sensitivity: 255 cis-pQTLs passed the significance threshold in the raw data matrix, and extra 133 cis-pQTLs have been detected after batch effect adjustment—100 after normalization and 33 additional cis-pQTLs after batch effect correction. On peptide level, this corresponds to 993 peptide-level pQTLs in raw data, additional 405 peptides after normalization and 352 after batch effect correction. Similarly, for the TMT mouse dataset 3,306 cis-pepQTLs were detected at LOD ≥ 4 among the 8,774 peptides with complete measurements in the normalized data, with an additional 109 cis-pepQTLs detected after correction for TMT batch factor. Note, that for the aging mouse dataset, normalization seems to alleviate most of the bias, as was already demonstrated in Fig EV1C. This effect is quite common and is the reason why the batch effect correction step was not performed in the InterLab study. Skipping batch effect correction is also reasonable, when the key question is "which proteins are differentially expressed", which is best address by ANOVA, or when it is not clear which technical factor should be corrected for (e.g., diagnostic plots do not show a clear clustering pattern and the weights of technical factors on PVCA are low). However, when biological groups are not confounded with technical factors, batch correction reduces noise and should improve the signal quality.

On the other hand, the sample and feature proximity methods we proposed in the "Quality control" section can be applied in almost any proteomic experimental context and thus are a more universal approach to quality control.

Sample proximity measured as correlation distributions for intra-batch vs. unrelated samples, and for replicates vs. all other samples can be seen for the aging mouse dataset pre- and post-adjustment in Figs 2B and 6B and for the Bariatric surgery study in Fig EV4C and D. This approach and considerations can also be applied with different distance measures and visualizations. Fig EV3G and H demonstrate this concept on the PanCancer data using Manhattan distance and a hierarchical clustering representation. These

Figure 6. Quality control of batch effect correction.

(A) Representative "Acads" protein QTL that in corrected data demonstrates a clear improvement in allele separation; (B) distribution of sample correlations between batches, within batches, and in replicated samples for corrected data, compare to Fig 2B; (C) heatmaps of peptide correlation for the proteins DHTKD1, GLO1, and HAAO, before and after correction: Correlation is positive for all peptides in the raw matrix, while after batch correction, the correlation of unrelated peptides becomes close to zero; (D) distribution of peptide correlation in raw data (brown) and in batch corrected data (yellow) for peptides from different proteins and peptides from the same protein: While same-protein peptide correlation is always higher than the correlation of unrelated peptides, the correlation of unrelated peptides approaches zero only after the correction.

proximity-based quality controls are universally applicable to any dataset with replicated samples and batches.

The second method is based on assessing the proximity of feature sets. In bottom-up proteomics, these are peptides or fragment ions. Similarly to the previous method, the distance between related peptides (i.e., belonging to a common protein) is expected to be smaller than the distance between unrelated ones. This effect is true not only for selected proteins (Fig 6C) but also holds true for the whole proteome (Fig 6D). The correlation of unrelated peptides, however, gets much closer to zero after the correction. Hence, before batch adjustment most of the peptide correlation was spurious.

In summary, through a combination of positive and negative quality controls, it is possible to assess improvements in data quality. It should, however, be stressed that quality control methods are not criteria for choosing batch adjustment methods. Their purpose is rather to provide metrics to control the quality of the data before and after adjustment.

Computational tools and analysis workflow for future applications

To facilitate the practical application of the steps and principles described here, we have developed an R package, named "proBatch" and made it available in Bioconductor (https://www.bioconductor.org/packages/release/bioc/html/proBatch.html). The "proBatch" package wraps multiple established techniques for data transformation and visualization with proven utility in the control of batch effects.

The package has been designed based on the following principles to facilitate the construction of batch effect adjustment pipelines:

For each step of the workflow, one or more functions are provided.

The input to each function is standardized to a feature measurement table (e.g., peptide, protein, fragment, transition) and a sample annotation table with technical and biological factors.

Consistent visual representation of factors at all steps of the analysis (e.g., color scheme defined once per analysis).

Note that we have also deliberately avoided the inclusion of a "single-click" batch adjustment function that would combine normalization and correction. As explained above, the choice of the specific algorithm for normalization and batch correction depends on the properties of a particular dataset, and thus, keeping these functions as separate exchangeable modules should prompt the researcher to check the properties of the data and the goals of the analysis and ultimately make an informed choice of appropriate batch adjustment methods.

To demonstrate how the package can be applied to different experimental setups, and how different parameters of the study can be adjusted, we make the code used for the analyses of the five case studies presented in this manuscript available as a GitHub repository (https://github.com/symbioticMe/batch_effects_workflow_code) of the workflow. An accompanying Docker container containing all the tools required to replicate the analyses is made available on DockerHub as a proBatch Docker container (https://hub.docker.com/r/digitalproteomes/probatch).

Conclusions

The meaningful analysis of data generated by large-scale proteomic studies, made possible by recent technological advances in data collection, is critically dependent on the statistical power required for systems biology and translational medicine studies. However, great power comes with batch effect baggage and requires specialized tools to handle this problem.

It is sometimes argued that batch effects should not be corrected, but rather incorporated into the downstream analysis. We agree that accounting for batch effects in differential expression analysis using ANOVA is backed up with decades of statistical research and practice. However, ANOVA has several important disadvantages. First, ANOVA is best suited to adjust for batch effects on fragment or peptide level, as this is the level where batch effects manifest, while in practice researchers are interested in protein-level protein expression differences; second, ANOVA adjusts for linear or discrete batch effects, while MS-based proteomics is often affected by non-linear signal drift; and finally, when using protein quantities outside of differential expression analyses, such as in protein–protein correlation comparison in molecular networks, adjusting for batch effects with classical statistic approaches can be difficult. This explains why most large-scale studies opt for adjusted batch effects and proceed with a "batch-free" dataset.

Another common advice is to not correct batch effects to avoid "flattening" the biological signal. This can occur when the biological groups are confounded with the batch factors. However, confounding is a serious experimental design error and can be prevented by balancing the biological groups between the expected batches. On the other hand, to account for ill-defined batches (e.g., MS batches) samples can be randomized. "Overcorrection'' can also potentially occur with small-sized batches, where fitting a non-linear curve or accurately estimating the mean and the variance becomes difficult. Hence, it has been suggested that batches should contain at least 25 samples (Alter *et al*, 2000; Benito *et al*, 2004). To check whether small batches add too much noise, one can test the stability of the

significant hits with and without the batch in question. In most cases, however, the strength of the signal is sufficiently assessed at the quality control step, described in detail in this manuscript.

While most batch effect methods are similar across "omics" fields, missing values are more typical in MS proteomics than transcriptomics or genomics data. Missing values are often batch-associated (Karpievitch *et al*, 2012) and can cause established clustering and correction methods to fail. For example, missing values often lead to "uncorrectable" batch effects seen as batch clusters in PCA. This, however, is often due to filling the batch-specific missing values with zeros or small random numbers. Other methods, such as hierarchical clustering or ComBat (Johnson *et al*, 2007), could potentially be adapted to account for the sparsity of proteomic data. Even so, proteomics could benefit from the development of more methods robust against missing data.

Specialized proteomic applications could also benefit from further batch adjustment methodological developments. Here, we presented a workflow and applied it to DIA and DDA (label free and TMT based) data. The workflow relies on having a quantitative data matrix with few assumptions on its structure and is applicable to most quantitative proteomic workflows. However, proteomic applications incorporating affinity purification, size exclusion chromatography, PTM-centric protocols, or subcellular fractionation are typically more heterogeneous. These dataset types will likely require additional efforts for batch effect characterization and correction.

While searching for publicly available datasets to present as case studies in this manuscript, we also noted that many of the available studies have been deposited with little to none of the technical meta-data required to perform batch correction. We believe that this limits the re-usability of proteomic data and would like to encourage future studies to include this information.

In conclusion, mass spectrometry-based proteomics has come a long way and is continuing to evolve. As throughput and reproducibility increase, so do batch effect-related issues. Hence, we expect experimental design and batch effect correction methods to also grow in importance and to take center stage in large-scale proteomic applications such as clinical proteomics and systems biology.

Expanded View for this article is available online.

Author contributions

JČ and PGAP designed the analysis workflow and wrote the manuscript; EGW introduced multiple components in the workflow and made major

contributions to the manuscript outline and text. JČ, CHL, EGW, VSS, and GRK performed data analysis. JČ, CHL, and PGAP implemented back-end R functionality; EGW helped with the code testing. EGW, TS, BCC, and GRK provided proteomic data and contributed to workflow conceptualization. FW, SG, and BW provided proteomic data measurements and analysis. MRM provided critical input on the project and assisted in analysis pipeline design. MRM and RA commented and edited the manuscript. PGAP and RA supervised the study.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Akey JM, Biswas S, Leek JT, Storey JD (2007) On the design and analysis of gene expression studies in human populations. *Nat Genet* 39: 807–809

Alter O, Brown PO, Botstein D (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci USA* 97: 10101–10106

Benito M, Parker J, Du Q, Wu J, Xiang D, Perou CM, Marron JS (2004) Adjustment of systematic microarray data biases. *Bioinforma Oxf Engl* 20: 105–114

Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19: 185–193

Brenes A, Hukelmann J, Bensaddek D, Lamond AI (2019) Multibatch TMT reveals false positives, batch effects and missing values. *Mol Cell Proteomics MCP* 18: 1967–1980

Chawade A, Alexandersson E, Levander F (2014) Normalyzer: a tool for rapid evaluation of normalization methods for omics data sets. *J Proteome Res* 13: 3114–3120

Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, Liu C (2011) Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS One* 6: e17238

Choi M, Chang C-Y, Clough T, Broudy D, Killeen T, MacLean B, Vitek O (2014) MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics* 30: 2524–2526

Clough T, Thaminy S, Ragg S, Aebersold R, Vitek O (2012) Statistical protein quantification and significance analysis in label-free LC-MS experiments with complex designs. *BMC Bioinformatics* 13(Suppl 1): S6

Collins BC, Hunter CL, Liu Y, Schilling B, Rosenberger G, Bader SL, Chan DW, Gibson BW, Gingras A-C, Held JM *et al* (2017) Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. *Nat Commun* 8: 291

Collins BC, Hunter CL, Liu Y, Schilling B, Rosenberger G, Bader SL, Chan DW, Gibson BW, Gingras AC, Held JM *et al* (2017) ProteomeXchange PXD004886 (http://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD004886). [DATASET]

Čuklina J, Pedrioli PGA, Aebersold R (2020) Review of batch effects prevention, diagnostics, and correction approaches. In *Mass spectrometry data analysis in proteomics*, Matthiesen R (ed.), pp 373–387. New York, NY: Springer New York

Dillies M-A, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J *et al* (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* 14: 671–683

Durbin BP, Hardin JS, Hawkins DM, Rocke DM (2002) A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics* 18: S105–S110

Ebhardt HA, Sabidó E, Hüttenhain R, Collins B, Aebersold R (2012) Range of protein detection by selected/multiple reaction monitoring mass spectrometry in an unfractionated human cell culture lysate. *Proteomics* 12: 1185–1193

Escher C, Reiter L, Maclean B, Ossola R, Herzog F, Chilton J, MacCoss MJ, Rinner O (2012) Using iRT, a normalized retention time for more targeted measurement of peptides. *Proteomics* 12: 1111–1121

Evans C, Hardin J, Stoebel DM (2018) Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief Bioinform* 19: 776–792

Gilad Y, Mizrahi-Man O (2015) A reanalysis of mouse ENCODE comparative gene expression data. *F1000Research* 4: 121

Goh WWB, Wang W, Wong L (2017) Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol* 35: 498–507

Gregori J, Villarreal L, Méndez O, Sánchez A, Baselga J, Villanueva J (2012) Batch effects correction improves the sensitivity of significance tests in spectral counting-based comparative discovery proteomics. *J Proteomics* 75: 3938–3951

Hu J, Coombes KR, Morris JS, Baggerly KA (2005) The importance of experimental design in proteomic mass spectrometry experiments: Some cautionary tales. *Brief Funct Genomic Proteomic* 3: 322–331

Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8: 118–127

Karpievitch YV, Dabney AR, Smith RD (2012) Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinformatics* 13: S5

Kaufman S, Arnold M, Diaz AA, Neubauer H, Wolfrum S, Köfeler H, Langhans W, Krieger J-P (2019) Roux-en-Y gastric bypass surgery reprograms enterocyte triglyceride metabolism and postprandial secretion in rats. *Mol Metab* 23: 51–59

Kaufman S, Arnold M, Diaz AA, Neubauer H, Wolfrum S, Köfeler H, Langhans W, Krieger J-P (2019) MassIVE MSV000087519 (https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?accession=MSV000087519) [DATASET]

Keele GR, Zhang T, Pham DT, Vincent M, Bell TA, Hock P, Shaw GD, Munger SC, de Villena FP-M, Ferris MT *et al* (2020) Regulation of protein abundance in genetically diverse mouse populations. *bioRxiv* https://doi.org/10.1101/2020.09.18.296657 [PREPRINT]

Keele GR, Zhang T, Pham DT, Vincent M, Bell TA, Hock P, Shaw GD, Munger SC, de Villena FP-M, Ferris MT *et al* (2020) ProteomeXchange PXD018886 (http://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD018886). [DATASET]

Landfors M, Philip P, Rydén P, Stenberg P (2011) Normalization of high dimensional genomics data where the distribution of the altered variables is skewed. *PLoS One* 6: e27942

Lazar C, Meganck S, Taminau J, Steenhoff D, Coletta A, Molter C, Weiss-Solis DY, Duque R, Bersini H, Nowe A *et al* (2013) Batch effect removal methods for microarray gene expression data integration: a survey. *Brief Bioinform* 14: 469–490

Lee AH, Shannon CP, Amenyogbe N, Bennike TB, Diray-Arce J, Idoko OT, Gill EE, Ben-Othman R, Pomat WS, van Haren SD *et al* (2019) Dynamic molecular changes during the first week of human life follow a robust developmental trajectory. *Nat Commun* 10: 1092

Leek JT, Scharpf RB, Bravo HCHC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* 11: 733–739

Liu Y, Buil A, Collins BC, Gillet LCJ, Blum LC, Cheng L-Y, Vitek O, Mouritsen J, Lachance G, Spector TD *et al* (2015) Quantitative variability of 342 plasma proteins in a human twin population. *Mol Syst Biol* 11: 786

Luan H, Ji F, Chen Y, Cai Z (2018) Quality control-based signal drift correction and interpretations of metabolomics/proteomics data using random forest regression. *bioRxiv* https://doi.org/10.1101/253583 [PREPRINT]

Luo J, Schumacher M, Scherer A, Sanoudou D, Megherbi D, Davison T, Shi T, Tong W, Shi L, Hong H *et al* (2010) A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. *Pharmacogenomics J* 10: 278–291

Matafora V, Corno A, Ciliberto A, Bachi A (2017) Missing value monitoring enhances the robustness in proteomics quantitation. *J Proteome Res* 16: 1719–1727

Mertins P, Mani Dr, Ruggles KV, Gillette MA, Clauser KR, Wang P, Wang X, Qiao JW, Cao S, Petralia F *et al* (2016) Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* 534: 55–62

Muntel J, Kirkpatrick J, Bruderer R, Huang T, Vitek O, Ori A, Reiter L (2019) Comparison of protein quantification in a complex background by DIA and TMT workflows with fixed instrument time. *J Proteome Res* 18: 1340–1351

Oberg AL, Vitek O (2009) Statistical design of quantitative mass spectrometry-based proteomic experiments. *J Proteome Res* 8: 2144–2156

Okada H, Ebhardt HA, Vonesch SC, Aebersold R, Hafen E (2016) Proteome-wide association studies identify biochemical modules associated with a wing-size phenotype in Drosophila melanogaster. *Nat Commun* 7: 12649

Peirce JL, Lu L, Gu J, Silver LM, Williams RW (2004) A new set of BXD recombinant inbred lines from advanced intercross populations in mice. *BMC Genet* 5: 7

Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC *et al* (2002) Use of proteomic patterns in serum to identify ovarian cancer. *Lancet Lond Engl* 359: 572–577

Rice JA (2006) *Mathematical statistics and data analysis*. Third, Belmont, CA: Duxbury Press

Rosenberger G, Ludwig C, Röst HL, Aebersold R, Malmström L (2014) aLFQ: An R-package for estimating absolute protein quantities from label-free LC-MS/MS proteomics data. *Bioinformatics* 30: 2511–2513

Röst HL, Liu Y, D'Agostino G, Zanella M, Navarro P, Rosenberger G, Collins BC, Gillet L, Testa G, Malmström L *et al* (2016) TRIC: An automated alignment strategy for reproducible protein quantification in targeted proteomics. *Nat Methods* 13: 777–783

Sajic T, Liu Y, Arvaniti E, Surinova S, Williams EG, Schiess R, Hüttenhain R, Sethi A, Pan S, Brentnall TA *et al* (2018) Similarities and differences of blood N-glycoproteins in five solid carcinomas at localized clinical stage analyzed by SWATH-MS. *Cell Rep* 23: 2819–2831

Sajic T, Liu Y, Arvaniti E, Surinova S, Williams EG, Schiess R, Hüttenhain R, Sethi A, Pan S, Brentnall TA *et al* (2018) ProteomeXchange PXD004998 (http://proteomecentral.proteomexchange.org/cgi/GetDataset?ID= PXD004998). [DATASET]

Schubert OT, Röst HL, Collins BC, Rosenberger G, Aebersold R (2017) Quantitative proteomics: challenges and opportunities in basic and applied research. *Nat Protoc* 12: 1289–1294

Shabalin AA, Tjelmeland H, Fan C, Perou CM, Nobel AB (2008) Merging two gene-expression studies via cross-platform normalization. *Bioinformatics* 24: 1154–1160

Shen X, Gong X, Cai Y, Guo Y, Tu J, Li H, Zhang T, Wang J, Xue F, Zhu ZJ (2016) Normalization and integration of large-scale metabolomics data using support vector regression. *Metabolomics* 12: 89

Teo G, Kim S, Tsou CC, Collins B, Gingras AC, Nesvizhskii AI, Choi H (2015) mapDIA: preprocessing and statistical analysis of quantitative proteomics data from data independent acquisition mass spectrometry. *J Proteomics* 129: 108–120.

Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, Mann M, Cox J (2016) The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods* 13: 731–740

Välikangas T, Suomi T, Elo LL (2018) A systematic evaluation of normalization methods in quantitative label-free proteomics. *Brief Bioinform* 19: 1–11

Wang M, Jiang L, Jian R, Chan JY, Liu Q, Snyder MP, Tang H (2021) RobNorm: model-based robust normalization method for labeled quantitative mass spectrometry proteomics data. *Bioinforma Oxf Engl* 37: 815–821

Williams EG, Pfister N, Roy S, Statzer C, Haverty J, Ingels J, Bohl C, Hasan M, Čuklina J, Bühlmann P *et al* (2021) Multi-omic profiling of the liver across diets and age in a diverse mouse population. *bioRxiv* https://doi.org/10. 1101/2020.08.20.222968 [PREPRINT]

Williams EG, Pfister N, Roy S, Statzer C, Haverty J, Ingels J, Bohl C, Hasan M, Čuklina J, Bühlmann P *et al* (2021) ProteomeXchange PXD009160 (http:// proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD009160). [DATASET]

Williams Eg, Wu Y, Jha P, Dubuis S, Blattmann P, Argmann Ca, Houten Sm, Amariuta T, Wolski W, Zamboni N *et al* (2016) Systems proteomics of liver mitochondria function. *Science* 352: aad0189

Zhang B, Wang J, Wang X, Zhu J, Liu Qi, Shi Z, Chambers MC, Zimmerman LJ, Shaddox KF, Kim S *et al* (2014) Proteogenomic characterization of human colon and rectal cancer. *Nature* 513: 382–387

Zhang H, Liu T, Zhang Z, Payne SH, Zhang B, McDermott JE, Zhou J-Y, Petyuk VA, Chen Li, Ray D *et al* (2016) Integrated proteogenomic characterization of human high-grade serous ovarian cancer. *Cell* 166: 755–765