Journal of proteome research

pubs.acs.org/jpr Article

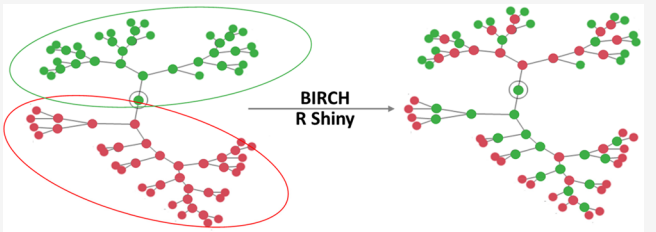
BIRCH: An Automated Workflow for Evaluation, Correction, and Visualization of Batch Effect in Bottom-Up Mass Spectrometry-Based Proteomics Data

Niveda Sundararaman,* Archana Bhat, Vidya Venkatraman, Aleksandra Binek, Zachary Dwight, Nethika R. Ariyasinghe, Sean Escopete, Sandy Y. Joung, Susan Cheng, Sarah J. Parker,+ Justyna Fert-Bober,+ and Jennifer E. Van Eyk*,+

Cite This: https://doi.org/10.1021/acs.jproteome.2c00671 Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Recent surges in large-scale mass spectrometry (MS)-based proteomics studies demand a concurrent rise in methods to facilitate reliable and reproducible data analysis. Quantification of proteins in MS analysis can be affected by variations in technical factors such as sample preparation and data acquisition conditions leading to batch effects, which adds to noise in the data set. This may in turn affect the effectiveness of any biological conclusions derived from the data. Here we present Batch-effect Identification, Representation, and Correction of Heterogeneous data (BIRCH), a workflow for analysis and correction of batch effect through an automated, versatile, and easy to use web-based tool with the goal of eliminating technical variation. BIRCH also supports diagnosis of the data to check for the presence of batch effects, feasibility of batch correction, and imputation to deal with missing values in the data set. To illustrate the relevance of the tool, we explore two case studies, including an iPSC-derived cell study and a Covid vaccine study to show different context-specific use cases. Ultimately this tool can be used as an extremely powerful approach for eliminating technical bias while retaining biological bias, toward understanding disease mechanisms and potential therapeutics.



KEYWORDS: *proteomics, mass spectrometry, batch correction, imputation*

INTRODUCTION

Mass spectrometry (MS)-based proteomics involves the identification and quantification of proteins which comprise a proteome or subproteome and are known to be the key regulators of a majority of the biological processes. This makes MS-proteomics a valuable technique for not only understanding physiological processes and disease conditions,[1,2] but also in drug and biomarker development.[1,3,4] Adding to the vital contributions of MS-proteomics to the field of medicine and biological discovery, the recent advancements in MS have now made it possible to achieve higher sample-throughput, with hundreds of samples processed in a single batch.[5,6]

For large-scale MS-proteomic projects, it often becomes necessary to combine data from multiple batches derived from sample preparation, reagent lots, laboratory conditions, personnel, and instrumentation.[7,8] This tends to introduce batch effects into the data sets, adding unnecessary noise to the data, reducing the statistical power of the analysis, and finally leading to inconsistent or unreliable biological inference.[9] Thus, to ensure that the sensitivity and reproducibility of the sample preparation and acquisition is being properly met, the evaluation of batch effects and correction thereafter becomes necessary before downstream analysis of MS-proteomics data.[9,10]

Several methods and algorithms have been established to remove systemic biases and increase the accuracy in the depiction of biological conclusions.[9,11,12] One such common method is normalization of the data, where the method of normalization is determined based on assumptions about the experiment design and the protein expression levels.[13−18] Some methods include normalization based on data distribution, quantile normalization, median normalization, and control-based normalization, all of which aid in transforming read or intensity counts to meaningful quantification of biological expression. However, in some cases, data normalization may not be sufficient in removing technical bias where overall distribution of samples may be normalized across batches, but

Special Issue: Software Tools and Resources 2023

Received: October 17, 2022

individual feature distribution could still be affected by batch-level bias. Thus, further data conditioning is needed in eliminating the remaining technical bias, and this is where the process labeled as batch-effect correction comes into effect.

The most popular algorithm for batch correction is comBat,[19] which performs further correction on a normalized raw data matrix. The comBat method of correction across batches has been applied to several existing bioinformatics tools in the field of genomics and most recently has been extended to proteomics focused tools such as proBatch,[10] BatchServer,[20] and CPTAC.[21] The proBatch and CPTAC tools, in addition to performing comBat correction, also detail normalization of the data set and log scaling. The proBatch tool also performs visualization comparing pre- and postcorrected data. However, both these tools do not provide a user-friendly web application for ease of use, nor do they perform any initial diagnosis to determine if the data set follows the rules necessary to enable correction. Additionally, filtering of missing data and data imputation is not available. BatchServer, on the other hand, does provide a user-friendly web application for batch correction along with data visualization, but an end-to-end pipeline following the steps of data diagnosis, missing value filtering, and extensive visualization highlighting the correction process is unavailable. Hence, there remains an outstanding need for a user-friendly tool that provides a systematic evaluation of normalization methods, facilitates imputation of missing data,[22,23] and provides visualization to assess differential factors contributing to batch effect. Here we present Batch-effect Identification, Representation, and Correction of Heterogeneous data (BIRCH), a user-friendly itemized workflow with easy-to-use web-based RShiny app to facilitate diagnosis, filtering, correction, and visualization of batch effects. Our tool (1) provides automated data processing and manipulation steps such as normalization, filtering, log scale transformation, batch-effect correction and imputation in an orderly fashion, along with identifying the best data-level to perform correction on, i.e., decision to correct on fragment, peptide or protein level, (2) aids in visualizing sample and feature level data distribution to see if patterns emerge across or within batches (for example: data missingness may be prominent in samples from a specific batch, indicating immediate causes to rerun the batch), (3) addresses the ambiguity in identifying if data is skewed by batch effect and if so, whether batch-effect correction is sufficient to eliminate the bias, and (4) provides extensive visualization and depiction of bias in the data set to identify the factors contributing to the most variation across the data, be it biological or technical factors.

Briefly, BIRCH allows users to upload data searched through commonly available algorithms such as OpenSWATH[24] or DIA-Neural networks (DIA-NN)[25] in order to evaluate the presence of batch effect and subsequently correct the effect with the aid of the R package, proBatch. Apart from this, certain considerations are required for creating the workflow for the batch correction process: (1) Data must be filtered based on missingness to ensure that correction can accurately be performed. Features with entirely missing batches will be problematic, since data points across each batch are used to estimate statistical measures within corrected data. (2) It is recommended that the data set to be corrected is at the lowest data level (ex: fragment level data in the case of OpenSWATH and precursor level in the case of DIA-NN), since protein level rolled up measures are typically derived from raw intensity measures at the fragment or precursor level.[26,27] (3) Data imputation is also another important step carried out after batch correction, primarily required for data visualization. This is done because imputing before correction will force correction to include imputed data and significantly affect the measures of intensity of the originally present data. (4) It is also recommended that data be log-scaled for the correction[10] and imputation[28] processes.

To illustrate the usage of BIRCH across different types of data, we present two case studies, both depicting Data Independent Acquisition (DIA) data searched using two different search algorithms: OpenSWATH and DIA-NN. Further usage of the corrected data in downstream analysis in order to make biological interpretations is also depicted.

METHODS

iPSC-Derived Cell Study

Samples. Total proteomic analysis was done using cell lysate from differentiated and undifferentiated induced pluripotent stem cells as well as primary cells. Cell lysates were solubilized in 5% sodium dodecyl sulfate (SDS) and 50 mM triethylammonium bicarbonate (TEAB) and sonicated to lyse samples and dissolve proteins. Samples were reduced using tris(2-carboxyethl) phosphine, and incubated at 37 °C for 15 min. Samples were alkylated using 20 mM methylmethanethiosulfonate and acidified using 2.5% phosphoric acid. Protein was trapped using S-Trap micro columns (Profiti) and washed using 100 mM TEAB in 90% methanol. Last, proteins were digested in 50 mM TEAB and incubated for 1 h at 37 °C on the S-Trap column. Eluted protein was dried and resuspended in 0.1% formic acid in water for injection on liquid chromatography (LC) MS/MS.

MS Data Acquisition. Data was acquired in data-independent acquisition (DIA) using a Thermo Fusion Lumos Orbitrap mass spectrometer (ThermoFisher, Bremen, Germany). Desalted peptides were separated on an Ultimate 3000 ultrahigh-pressure chromatography system with a 120 min gradient from 0 to 38% acetonitrile on a C18 column (15 cm length, 300 μm diameter) at a flow rate of 1.2 μL/min. Source parameters included spray voltage at 3 kV, capillary temp of 300 °C, and RF funnel level of 40%. MS1 resolutions were set to 60,000 and with the AGC target of 4 × 105 and maximum injection time of 50 ms. MS2 were sampled across a precursor mass range of 400−1000 and fragmented using HCD with collision energy of 32% with data collected at a 15,000 resolution and a maximum injection time of 30 ms and AGC target of 5 × 104. All data was acquired in profile mode using positive polarity.

Data Analysis. Data was analyzed using the DIA-NN software (version 1.8.0) using an in silico digested human reviewed and canonical FASTA library downloaded from the UniProt database (December 2020). RT-dependent cross run normalization was enabled. Library generation was set to "FASTA digest for library-free search" and MBR was enabled. When reporting protein numbers and quantities, the Protein.Group column in the DIA-NN's report was used to identify the protein group and the PG.Normalized column was used to obtain the normalized quantity. Next, the Protein.Group column was used to only include unique proteins that are quantified using proteotypic peptides only. Additionally, the missed cleavages rate was set to 2 and peptide length range was set between 5 and 30. All other settings were left default. The software output was filtered at precursor q-value <1%, and the global protein q-value <1% filter was also applied to all benchmarks.

https://doi.org/10.1021/acs.jproteome.2c00671
J. Proteome Res. XXXX, XXX, XXX−XXX

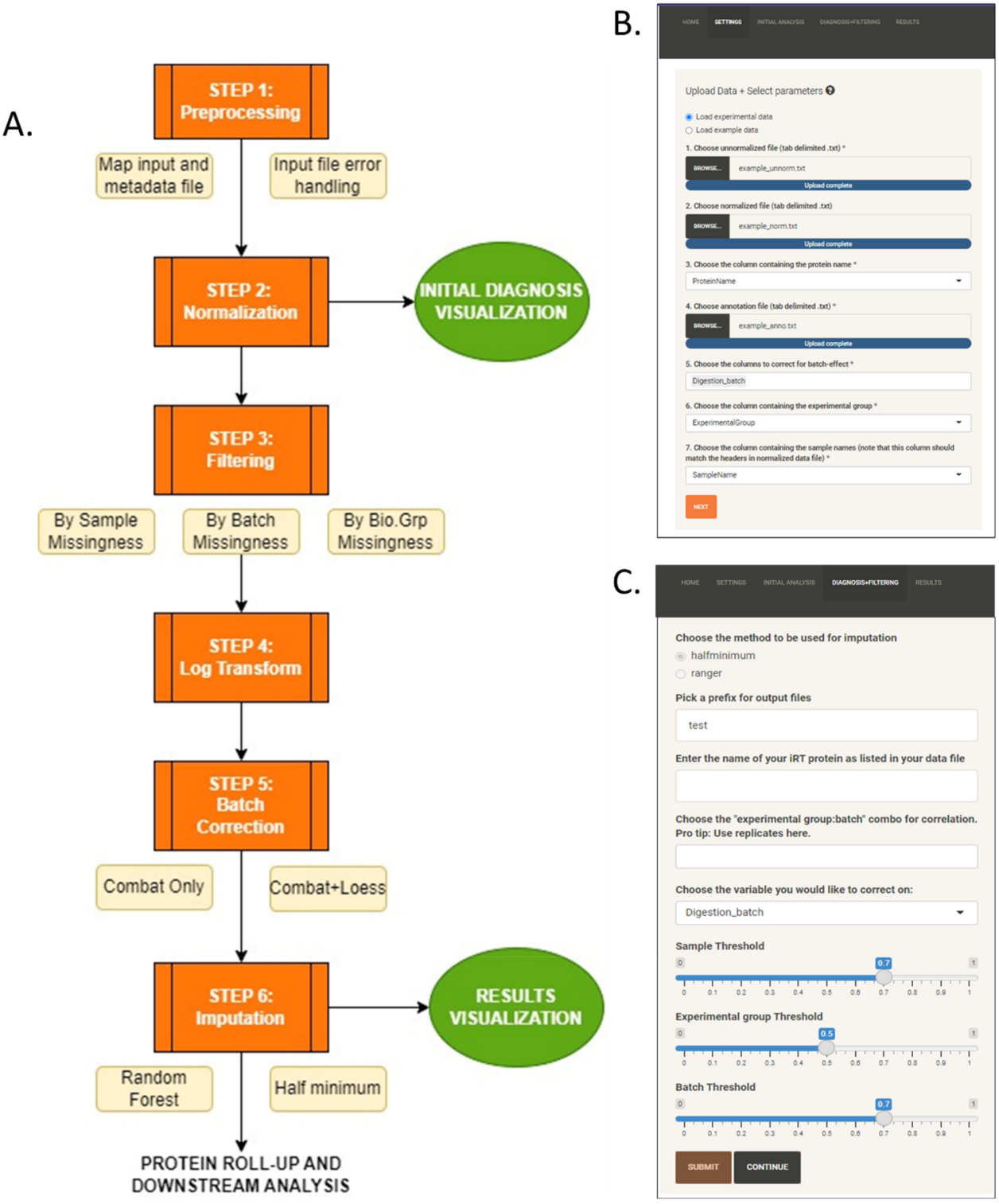


Figure 1. BIRCH web application. (A) BIRCH workflow. (B) Screenshot illustrating the Settings tab for easy upload of protein intensity data and sample information file, along with additional input parameters required from the user for appropriate batch correction and analysis. (C) Screenshot of the Diagnosis+Filtering tab to determine which factor needs batch correction. In this tab the user can also provide optional arguments to explore and analyze spiked-in iRT peptides and correlation between technical replicates. Lastly, filtering criteria can also be applied at this step to remove samples and fragments based on missingness levels.

SARS-CoV-2 Vaccine Study

Samples. The plasma samples were collected from a cohort of healthcare workers (N = 120, age 44 ± 13 years, 79% female, 53% white) who received Pfizer−BioNTech vaccination at Cedars-Sinai Medical Center.[29] Protein's quantification was measured at four time points: before or up to 3 d after dose 1; within 7−21 d after dose 1; within 7−21 d after dose 2; and within 8-weeks post dose 2 of BNT162b2 vaccination.

Automated Protein Digestion. The plasma samples were loaded into a deep 96-well titer plate (Beckman Coulter) with

https://doi.org/10.1021/acs.jproteome.2c00671
J. Proteome Res. XXXX, XXX, XXX−XXX

single- or multichannel pipettes (Eppendorf). The samples underwent protein denaturation, reduction, alkylation, and digestion using the Beckman i7 automated workstation (BeckmanCoulter) programmed for uniform mixing as previously described at a controlled temperature and modified with online automated desalting.[30] Briefly, proteins were denatured in a solution of 35% v/v 2-2-2 trifluoroethanol (TFE, Sigma), 40 mM dithiothreitol (DTT, Sigma), and dissolved in 50 mM NH_4CO_3 (Sigma). The sample was denatured for 1 h at 60 °C. Samples were then alkylated for 30 min at 25 °C in the dark with the addition of iodoacetamide (Sigma, (10 mM final concentration). To prevent overalkylation DTT was added at final concentration of 5 mM and samples were incubated for a further 15 min at 25 °C. Next, a volume of 50 mM NH4CO3 was added to dilute out TFE to a final conc of 5 or 10%. Trypsin was added to a ratio of 25:1, and samples were incubated for 4 or 16 h at 37 or 42 °C. Digestion reactions were quenched with 5 μL of 25% FA. The plates were centrifuged at 3400 rpm for 5 min at 4 °C. Desalting was carried out using a positive pressure apparatus (Amplius Positive Pressure ALP, Beckman Coulter).

MS Acquisition. DIA analysis was performed on an Orbitrap Exploris 480 (Thermo) instrument interfaced with a flex source coupled to an Ultimate 3000 ultrahigh-pressure chromatography system with 0.1% mL/L formic acid in water as mobile phase A and 0.1 mL/L% formic acid in acetonitrile as mobile phase B. Peptides were separated on a linear gradient of 1−27% B organic phase for 45 min, 27−44% B for 15 min on a C18 column (15 cm with 300 μm ID, 3 μm Omega Polar C18 beads, and 100 Å pore size, Phenomenex) over the course of total 60 min at a flow rate of 9.5 μL/min. Between every sample the column was washed with a 10 min blank where the organic phase was increased to 98% and then re-equilibrated at 1% B for 2 min. Source parameters included spray at 3000 kV, capillary temp of 300 °C, and an RF funnel level of 40. MS1 resolution was set to 60,000 and AGC was set to "standard" with ion transmission of 100 ms. The mass range of 400−1000 and the AGC target value for fragment spectra of 300% were used. Peptide ions were fragmented at a normalized collision energy of 30%. Fragmented ions were detected across 50 DIA nonoverlapping 12 Da precursor windows. MS2 resolution was set to 15,000 with an ion transmission time of 25 ms. All data was acquired in profile mode using positive polarity.

Data Analysis. Raw intensity data for peptide fragments was extracted from DIA files using the open source openSWATH workflow[24] against the plasma library of the human twin population (February 2015) peptide assay library. Target and decoy peptides were then extracted, scored, and analyzed using the mProphet algorithm[31] to determine scoring cut-offs consistent with a 1% false discovery rate (FDR). Peak group extraction data from each DIA file was combined using the "feature alignment" script, which performs data alignment and modeling analysis across an experimental data set,[32] and transition-level data was normalized by MS2TIC.

Software Design and Availability

BIRCH is a tool designed to access and correct batch effects in bottom-up proteomics data in order to strengthen biological inferences free from technical bias. To utilize these features in BIRCH, the sample matrix with intensities at the fragment or precursor-level should be provided in the form of a tab delimited text (TXT) document. An unnormalized sample matrix file is mandatory, in which case normalization will be performed by BIRCH, but a prenormalized sample matrix file can also be provided. Additionally, a sample information file must be provided as a tab delimited text (TXT) file, with columns indicating the sample names (corresponding to that used in the sample matrix file), technical factors that could potentially cause the batch effect (for example, sample digestion batches or MS acquisition batches), and the biological group whose variation should be retained even after batch-correction.

The features of BIRCH are described in more detail in the Results and Discussion section. BIRCH is programmed using R 3.6.3, and the web-application for BIRCH is created using the R shiny framework. The source code for BIRCH, the link to the web application, and a usage documentation along with a tutorial using an example data set are included in the public GitHub repository, https://github.com/csmc-vaneykjlab/birch.

RESULTS AND DISCUSSION

Overview and Features

BIRCH is a web-based tool that allows users to perform batch correction and visualization on protein data. To do so, the following steps are executed: (1) preprocessing of input data, (2) data normalization, (3) filtering of samples and features, (4) log transformation, (5) batch correction using the R-based tool, proBatch, (6) imputation, and (7) visualization of batch effects (Figure 1A).

The preprocessing step reads, processes, and stores the DIA-MS processed data file containing the raw intensity data of peptide fragments, and maps it with the sample information file containing metadata columns including biological effects and potential batch effects for each sample. Considerations are made to ensure that the input data is valid (i.e., samples in the data file maps to the sample information file), readily available (i.e., raw intensity data is present and is numeric), and nonredundant. This is followed by the optional step, quantile normalization. This step can be skipped if the input data is already normalized by other methods. Next, the data is filtered to drop (a) samples having missing values over a set threshold and (b) fragments having missing values over a set threshold by batch and by biological groups. Filtered data is then transformed to the log2 scale before being batch corrected.

BIRCH provides the option to choose the batch effect procedure based on what would be appropriate for the biological background and data properties, especially those detected through the previous steps. The correction step attempts to remove two types of effects: continuous and discrete effects. Continuous effects, which could be manifested as MS signal drifts within each batch, are corrected using LOESS fitting[33] whereas discrete effects, which are manifested as feature-specific drifts across each batch, are corrected using comBat. The user can choose between either comBat only correction or both comBat and LOESS-based correction for further downstream analysis. The next step performs imputation of missing values in the input data matrix, with an option of choosing between random forest (RF) imputation using the R package ranger[34] and half-minimum imputation.[35] The default recommended method of imputation is RF. Finally, BIRCH enables the visualization of batch effects before and after batch correction to evaluate the outcome of the correction.

Hereby, BIRCH provides a unique automation solution for batch correction and imputation, while simultaneously integrating visualization charts to observe the effects of technical factors across batches and to assess the extent to which batch correction

https://doi.org/10.1021/acs.jproteome.2c00671
J. Proteome Res. XXXX, XXX, XXX−XXX

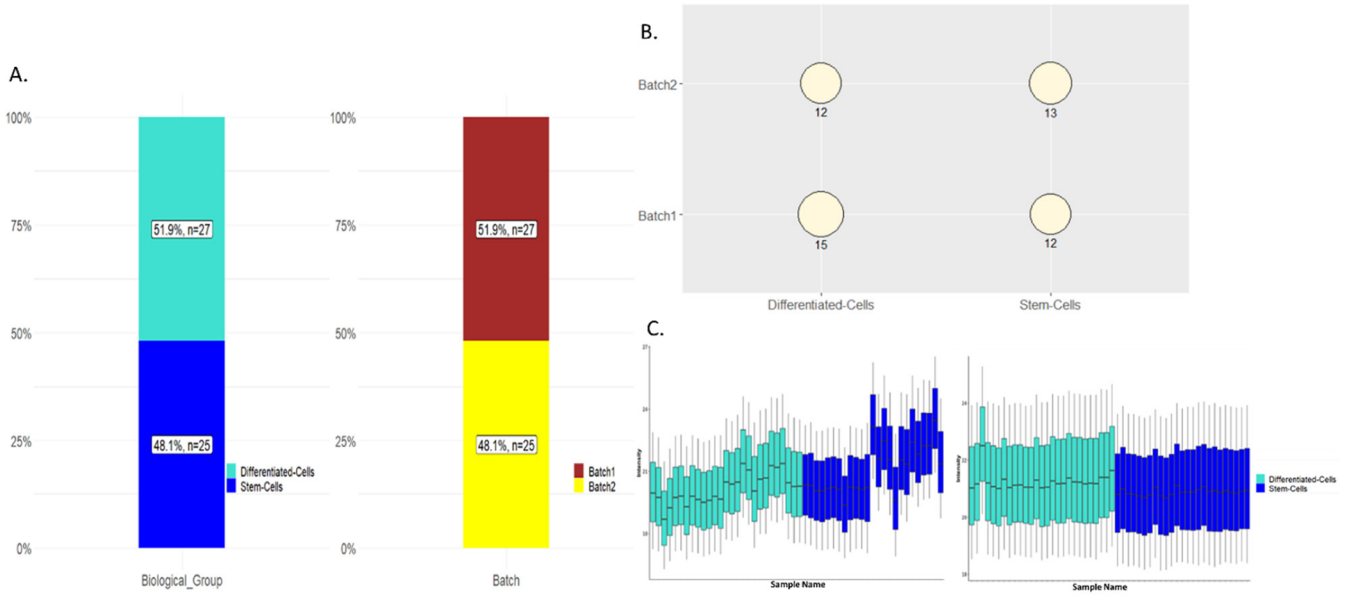


Figure 2. Distribution of samples across the biological group (cell types) and MS batch, with impact of normalization on the data. (A) Stacked bar-plot showing number and percentage of samples belonging to each the cell type and MS batch. In this case, sample distribution shows almost equal distribution in sample numbers across the two different cell types and batches. (B) Matrix showing sample distribution across the biological group and MS batch simultaneously. This helps ensure a balanced distribution of cell types across the different batches, which is necessary for accurate batch correction and to maintain biological variation after batch correction. (C) Box plots showing the effect of TIC normalization on the protein intensities per sample. Left panel shows unnormalized data while the right panel shows normalized data.

helps eliminate these effects while retaining differences due to biological factors.

BIRCH Interface

To provide an easy-to-use web interface, the architecture of BIRCH is designed into 5 components: (1) Home page, (2) Parameter Settings, (3) Initial Analysis, (4) Diagnosis and Filtering, and (5) Results and Visualization.

The first module is the "Home" tab with a brief description of the tool features and a guideline for running the tool. Next, the Settings tab (Figure 1B) enables users to upload their data files including an unnormalized file, a normalized file, and a sample information file in TXT format containing columns indicating a sample-wise distribution of biological and technical factors that cause variation in the data set. Additionally, if a normalized file is not provided, quantile normalization is done on the unnormalized data. Once the input parameters are submitted, the interface is redirected to the Initial Analysis tab which assesses the potential effectiveness of batch correction based on sample distribution across batches and missingness pattern of the data set. If validity of performing batch correction is met, the next step is Diagnosis and Filtering (Figure 1C). This tab enables users to customize their analysis with several mandatory and optional parameters. Mandatory parameters include choice of imputation method (chosen between RF and half-minimum methods), technical factor to correct on, and sample and feature-level minimum observations filtering thresholds. Optional parameters include identifiers for plotting individual feature diagnostics (example plots for iRT peptides) and for plotting correlation charts of quality control QC samples to attest for drifts across batches. Once the run is complete, the Results and Visualization tab becomes active which creates several charts including normalization plots, principal component analysis (PCA), and principal variance components analysis (PVCA) to compare batch effect drifts before and after correction. Options to download the visualization report in the HTML format and result files after correction and imputation is supported by the user interface (UI).

Case Study 1: Initial Analysis and Batch Correction to Aid in Comparison of Stem Cells vs Differentiated Cells

Identification of batch effects to uncover if correction is necessary can be challenging. Several factors including even and randomized distribution of samples across batches and across the different biological conditions need to be considered before deciding on batch correction of a data set. To address this, we performed initial analysis and diagnosis for the data set containing two distinct types of cells to better understand the distribution and to estimate if the underlying data requirements are met to proceed with correction. Using BIRCH, we first extracted visual depictions of the data based on the normalized precursor-level data generated using DIA-NN software and the sample information file. The resulting plots showed that (1) there were 52 samples distributed across the two cell types, i.e., stem cells and differentiated cells and processed in two digestion and MS batches, (2) there was even distribution of the number of samples across the batches and each batch consisted of at least 25 samples, which was found to increase accuracy of correction,[16] and (3) the two cell types were evenly distributed across the batches (Figure 2A,B). Thus, initial analysis highlighted that batch correction can be performed since randomness in sample distribution and an even spread of samples across batches are key to ensuring appropriate batch-effect analysis and correction.

The next step was to diagnose the presence of batch effects and the need for batch correction. First, we ensure that data normalization was effective in harmonizing the data set across all samples (Figure 2C). The data was then visualized and represented to check for any further batch effects using (1) PCA which showed batches 1 and 2 clustering separately from one another (Figure 3), and (2) PVCA which showed that the highest variation in the data set came from technical factors, i.e.,

https://doi.org/10.1021/acs.jproteome.2c00671
J. Proteome Res. XXXX, XXX, XXX−XXX

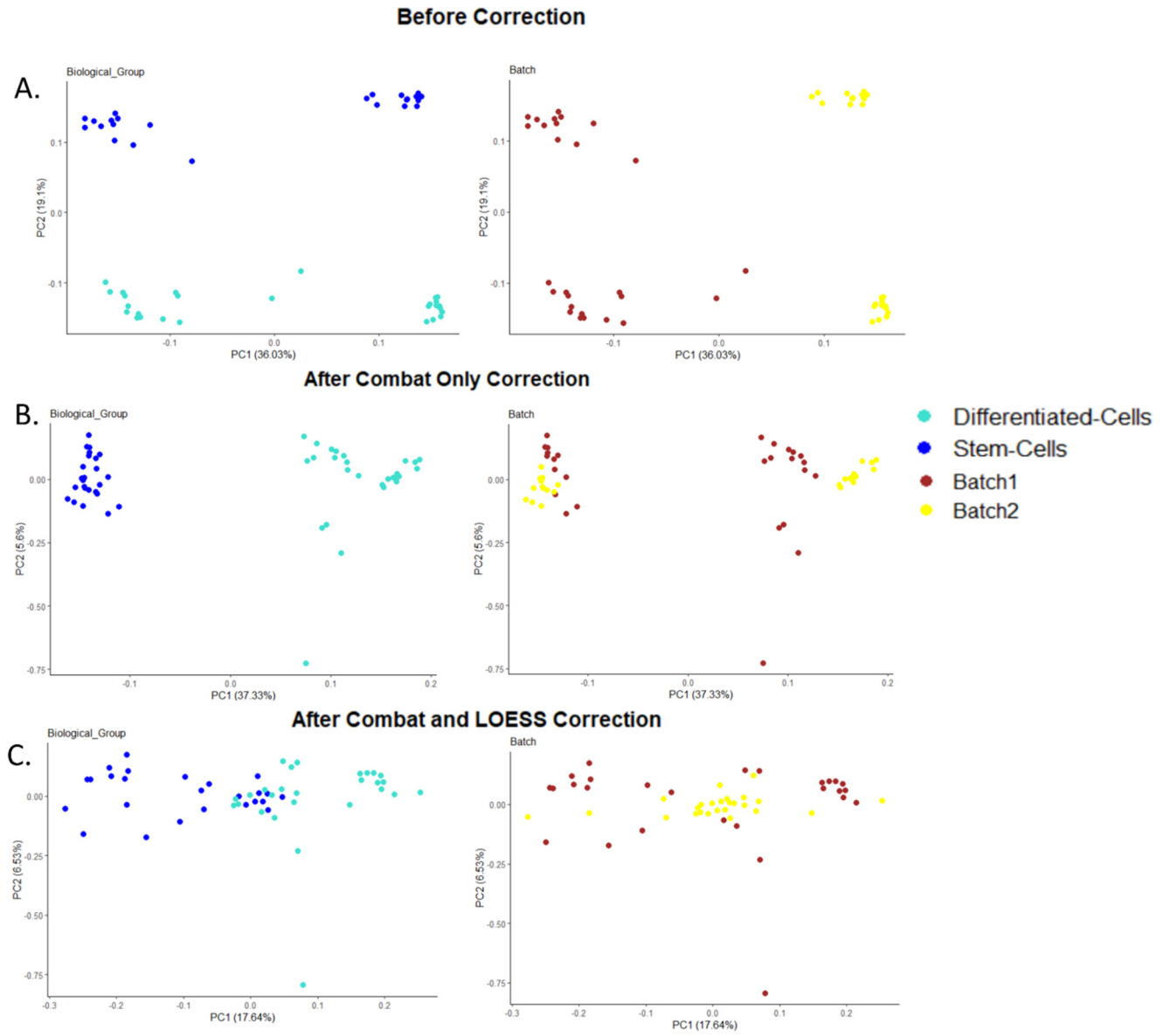


Figure 3. PCA before correction, after comBat only correction and after comBat + LOESS correction. Clusters are colored by cell type and batch. (A) Before correction, the PCA shows distinct clusters based on cell type as well as batch. (B) For after comBat only correction, the distinction due to cell type remains whereas distinction due to batch is corrected. (C) After comBat and LOESS correction, distinct clusters are no longer formed based on either cell type or batch, indicating that the data is overcorrected. The best correction method is the comBat only method.

the digestion and MS batch (Figure 4). Both the PCA and PVCA also revealed that while batch effects did exist, differences by cell type also had a role to play in contributing to variation in the data. Therefore, it is imperative that differences from biological variation are preserved, while attempting to eliminate the technical or analytical differences.

Based on these findings, the data was corrected using both comBat as well as comBat and LOESS methods, followed by RF imputation for further visualization. BIRCH has automated the decision process to evaluate the best method for correction based on (1) elimination of technical factors leading to variation in the data while simultaneously (2) retaining the variation caused by biological factors. ComBat only correction showed the elimination of batch effects arising from technical factors, while retaining the variation between the two cell types. On the other hand, comBat with LOESS correction attempted to eliminate effects from both technical as well as biological factors, causing overcorrection of the data. Hence, in this case the best method of correction is comBat only correction. This can then be used for further downstream analysis, including peptide and protein-level roll up, comparison of protein expression between the two cell types using tools such as mapDIA,[27] MSStats,[36] or MetaboAnalyst,[37] as well as Gene Ontology (GO) Term[38,39] and pathway analysis of top differentially expressed proteins (DEPs) using tools such as KEGG,[40] Reactome,[41] and WikiPathways.[42]

Case Study 2: Correction and Visualization of SARS-CoV-2 Vaccine Data

In this study, plasma samples were extracted and processed at four different time points before and after administering the SARS-CoV-2 vaccine, with the goal of deciphering the biological effect after vaccination in healthcare workers at Cedars-Sinai Medical Center. The batch correction was performed on the fragment-level using data generated through the OpenSWATH

https://doi.org/10.1021/acs.jproteome.2c00671
J. Proteome Res. XXXX, XXX, XXX−XXX

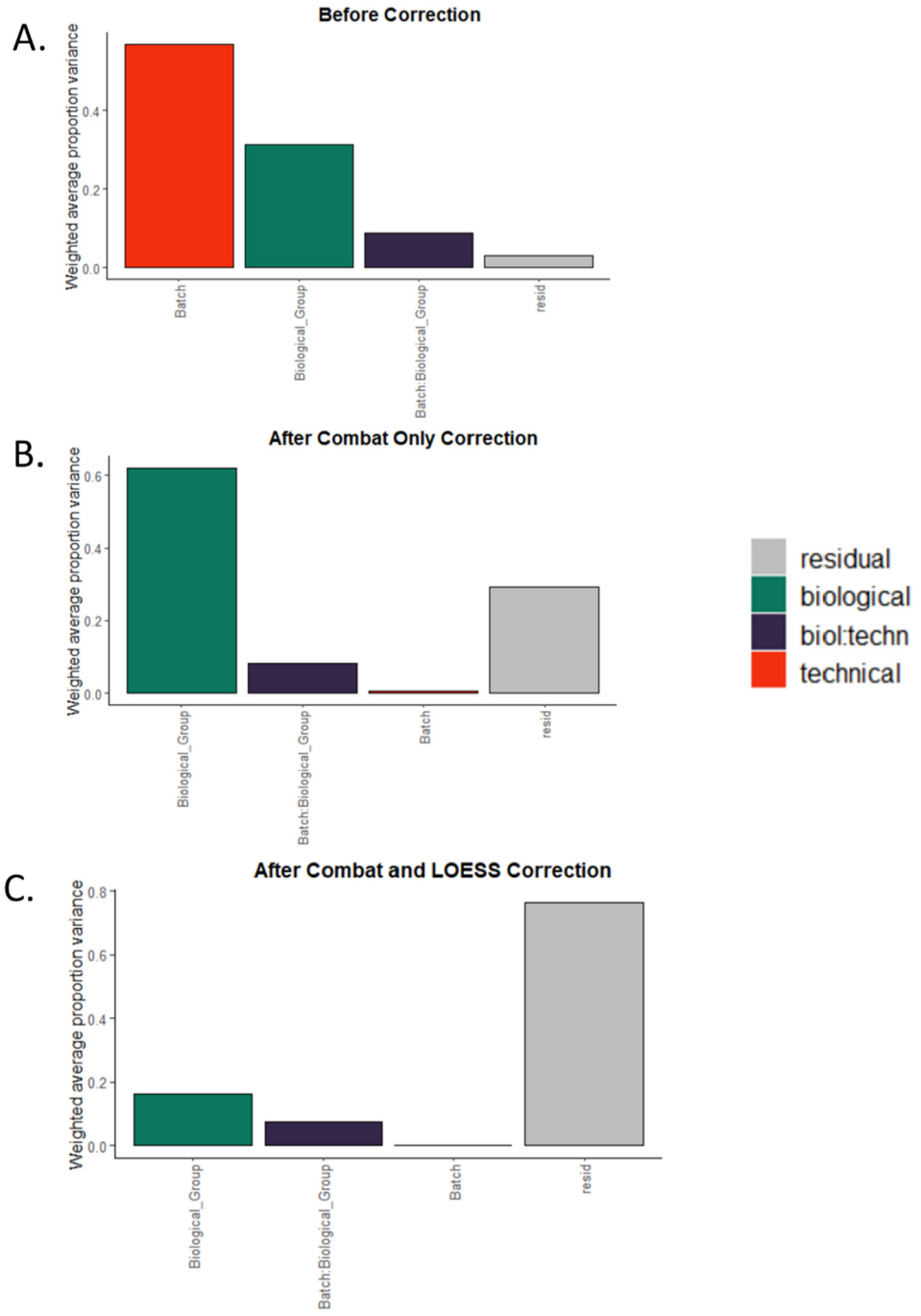


Figure 4. PVCA before correction, after comBat only correction, and after comBat + LOESS correction. Bars are colored by source of variation. (A) Before correction technical variation is the leading contributor to variance in the data set, hence correction is required. (B) After comBat-only correction seems to sufficiently eliminate most of the technical variation while retaining biological variation. (C) After comBat and LOESS correction tends to overcorrect the data, showing that variation by biological factor significantly drops whereas variation by technical factor remains the same. In all cases, a combined effect of batch and biological group remained as a source of variation. However, this remained consistent and showed a relatively smaller effect than the individual factors, thus ensuring that it did not affect with the correction process.

pipeline. Initial analysis revealed that there were 547 samples including the four different time points along with digestion and technical replicate QC samples spanning across 7 digestion plates and 13 corresponding MS batches. The assessment also proved that there was a uniform distribution in sample number across batches along with a balance between different time points across batches, indicating that data is eligible for correction in case of batch effects. The large data set showed irregular patterns of missingness distribution, invoking the need for data filtering to discard samples with ≥80% missingness and fragments with ≥50% missingness across all time points (Figure 5).

Next, the filtered data set was used to diagnose the presence of batch effect and the need for batch correction. Both the PCA and PVCA before correction proved the existence of batch effect: (1) PCA showed two distinct clusters separated by digestion and MS batches (Figure 6A), and (2) PVCA showed that digestion process was the most significant contributor to the data

https://doi.org/10.1021/acs.jproteome.2c00671
J. Proteome Res. XXXX, XXX, XXX−XXX

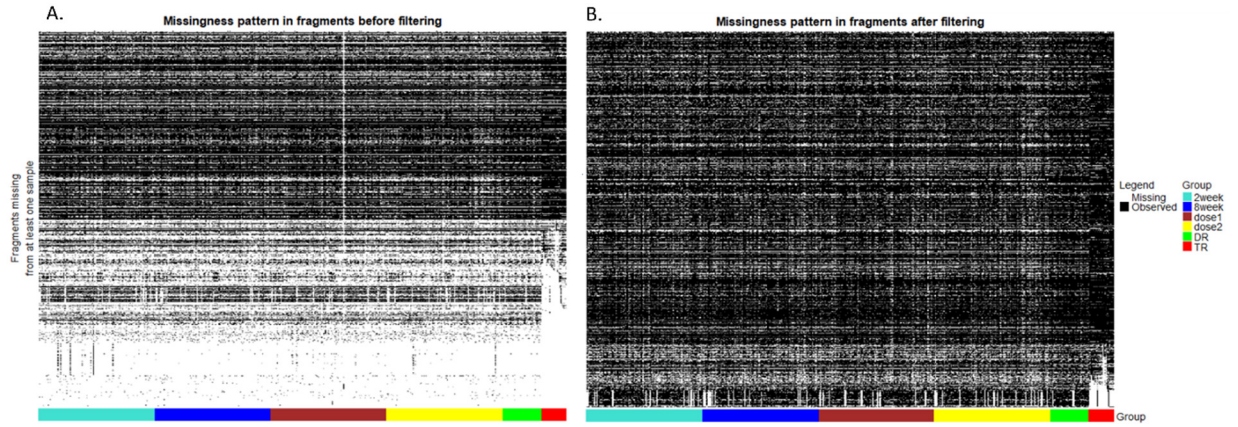


Figure 5. Missingness distribution across fragments before and after filtering. (A) Missingness distribution across fragments before filtering. (B) Missingness distribution across fragments after applying sample filtering threshold of ≥80% missingness and fragments filtering threshold of ≥50% missingness across all time points.

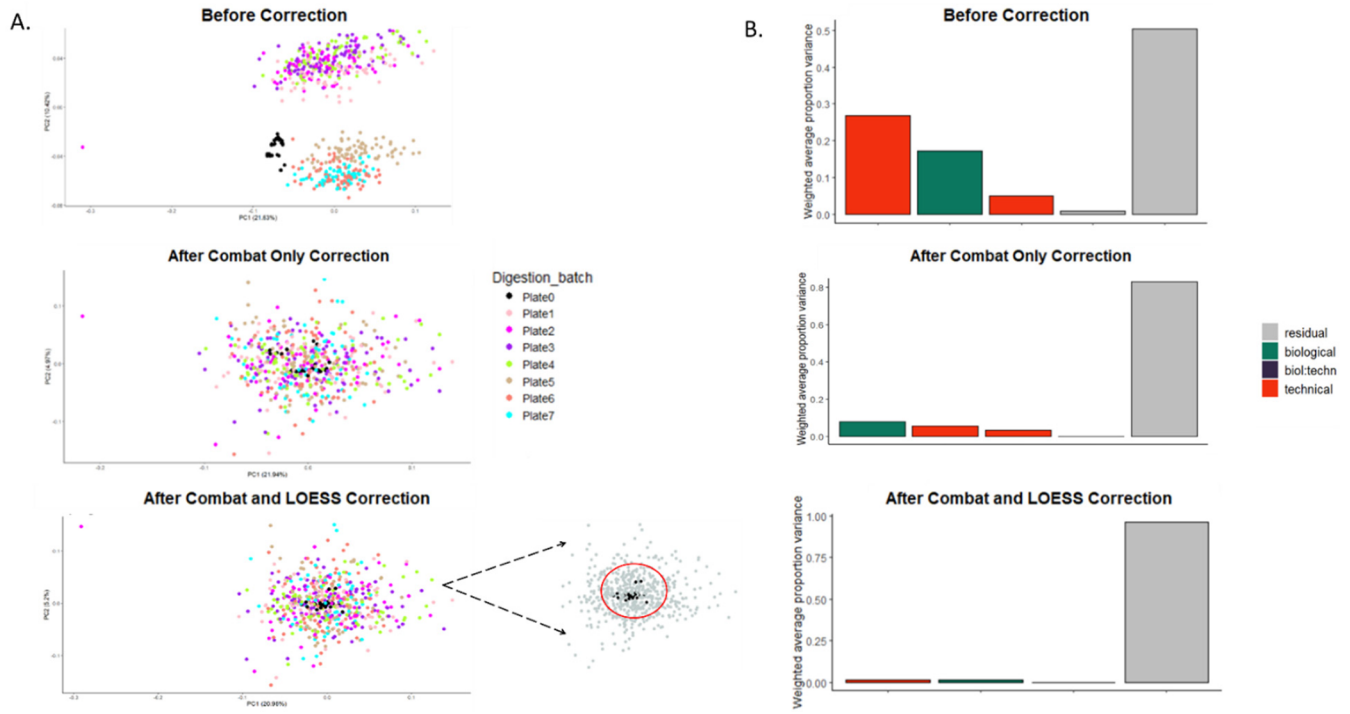


Figure 6. PCA and PVCA before and after correction. (A) PCA before correction shows distinct clustering between digestion plates 1−4 and digestion plates 5−7 respectively, indicating clear batch effects. PCA after comBat only correction eliminates the clustering by digestion batch, but variation within Plate 0 (indicating technical replicates added to monitor the correction methods) still exists. For PCA after comBat + LOESS batch correction performs the most appropriate correction where clustering by digestion batch is eliminated while ensuring that technical replicates retain the clustering as before correction (zoomed to show clustering of technical replicates). (B) PVCA before correction shows digestion batch having the highest variation. After comBat only correction partially removed technical variations while comBat + LOESS correction shows variation by digestion batch as almost completely eliminated.

variability, more so than variation from biological effects arising from the various time points (Figure 6B). Correction on digestion batch by comBat only and comBat with LOESS methods were performed along with RF imputation, with variation from technical factors dropping through the former and almost fully erased through the latter method.

To further verify the effectiveness of the correction on the digestion batch, we used the digestion QC samples to confirm the presence of batch-specific patterns at the fragment-level before correction, i.e., QC samples within a digestion plate showed high correlation whereas samples across plates showed lower correlation (Figure 7A). Upon correction, this trend no longer held true, with all samples within and across plates showing high correlation to one another. Additionally, for this analysis, indexed Retention Time (iRT) peptides, a standard set of 11 peptides from which normalized retention time can be derived,[43] were spiked in along with the samples. These were monitored to assess the manifestation of batch effects through feature-level drift in MS signal across the digestion plates (Figure 7B). In harmony with the previous examples, the feature-level also showed batch effects due to digestion, which was then

https://doi.org/10.1021/acs.jproteome.2c00671
J. Proteome Res. XXXX, XXX, XXX−XXX

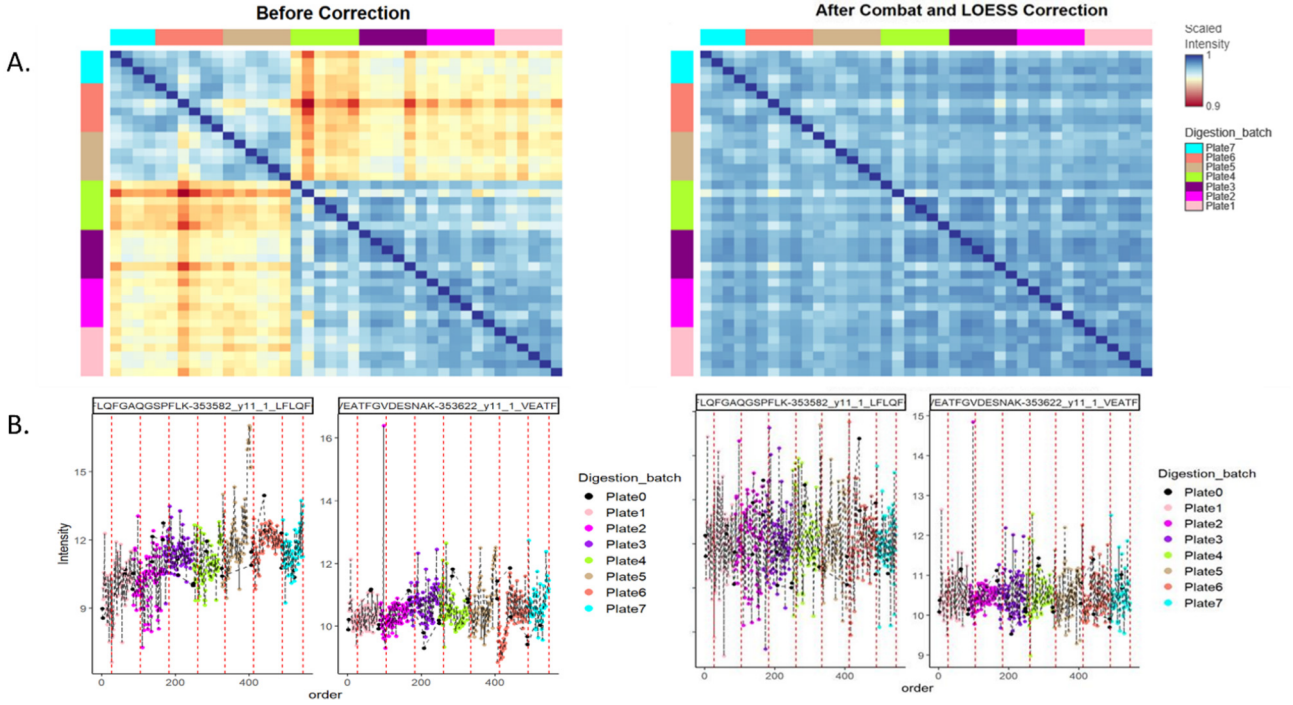


Figure 7. Correlation across digestion QC samples and feature-level MS drift. (A) Heatmaps showing correlation of digestion QC samples before and after comBat + LOESS correction. Before correction, the samples show high correlation within digestion plates, whereas after comBat + LOESS correction, correlation is equalized across plates. (B) iRT plots before and after comBat + LOESS correction shows that the MS signal drift at the feature level also reduced after correction.

resolved following correction by the comBat with LOESS method.

The corrected and imputed data generated by BIRCH was then processed using the mapDIA software to roll up fragment-level data to peptide and protein levels as well as to perform pairwise comparisons between samples at each time point against the dose 1 time point at the protein level. The result showed no statistically significant DEPs meeting the criteria of Bayesian FDR $\leq$0.05 with $\geq$1.5-fold change, indicating that the biological variation between individuals at every time point of taking the Covid vaccine is negligible.

CONCLUSIONS

We have developed a fully automated, easy-to-use bioinformatics tool called BIRCH to aid researchers in applying batch correction for different levels of MS-proteomic data, specifically DIA-MS data, including fragment, precursor, peptide, and protein levels as well as data sets generated through different software. BIRCH performs diagnosis, correction, imputation, and visualization of batch effects in order to eliminate technical variation and allow for meaningful biological conclusions devoid of bias. Using the two case studies, we have highlighted the importance of uniform and randomized sample distribution ensuring that biological groups are balanced across the different batches as well as the need to avoid small batch sizes to prevent overcorrection. Additionally, we have also demonstrated that simply performing normalization or QC monitoring of the digestion and acquisition processes does not ensure the absence of batch effects, and further correction may still be required. We acknowledge that there are certain areas for improvement within BIRCH; for example, we could increase options for performing different types of normalization based on data type beyond the currently supported option of quantile normalization. Another interesting addition would be to assess patterns in missingness to identify the best possible method of imputation. For example, it is suggested that for data sets having missingness at random (MAR), LLS, and RF imputation methods perform best whereas for data sets having missingness not at random (MNAR), left-censored imputation methods such as LOD and ND are ideal.[44]

There are several limitations to the current platform. First, the need for uniform and randomized distribution as highlighted by the initial diagnosis in both case studies shown here may limit the ability of the tool in scenarios where diverse data is being integrated from longitudinal studies or multiple sources, since it is more likely that the batch factors will be confounding with other factors of interest. Second, the platform's default of the RF imputation process can be very time-consuming depending on the size of the data sets. Improving the performance of the imputation method along with selection of optimal imputation methods for a data set as carried out by NAGuideR[45] by incorporation of a mechanism for evaluation of missing values would be a beneficial addition. A third limitation is that the case studies introduced here are both DIA-MS data sets. Expanding the use of BIRCH to label-free and label-based DDA, as showcased in other studies that also use the comBat correction technique,[10] would increase its usability in this field. In the future, we also plan to use BIRCH for batch correction of highly complex data sets such as those derived from single cell proteomics experiments where several hundred proteins can be quantified across thousands of cells in a single MS-proteomic project.[46]

In summary, BIRCH provides a versatile, user-friendly, automated solution for batch correction and imputation, specialized to handle bottom-up proteomics data. BIRCH can seamlessly process data sets acquired through various instruments and searches using different software, while providing a

https://doi.org/10.1021/acs.jproteome.2c00671
J. Proteome Res. XXXX, XXX, XXX−XXX

variety of customization options to enable normalization, filtering, correction, and imputation of the data. Ultimately, BIRCH can serve as an extremely powerful tool that ensures elimination of technical bias toward better understanding biological conclusions.

AUTHOR INFORMATION

Corresponding Authors

Niveda Sundararaman − *Smidt Heart Institute, Cedars-Sinai Medical Center, Los Angeles, California 90048, United States; Advanced Clinical Biosystems Research Institute, Cedars-Sinai Medical Center, Los Angeles, California 90048, United States;* orcid.org/0000-0001-7708-3754; Email: Niveda.Sundararaman@cshs.org

Jennifer E. Van Eyk − *Smidt Heart Institute, Cedars-Sinai Medical Center, Los Angeles, California 90048, United States; Advanced Clinical Biosystems Research Institute, Cedars-Sinai Medical Center, Los Angeles, California 90048, United States;* orcid.org/0000-0001-9050-148X; Email: Jennifer.VanEyk@cshs.org

Authors

Archana Bhat − *Smidt Heart Institute, Cedars-Sinai Medical Center, Los Angeles, California 90048, United States; Advanced Clinical Biosystems Research Institute, Cedars-Sinai Medical Center, Los Angeles, California 90048, United States*

Vidya Venkatraman − *Smidt Heart Institute, Cedars-Sinai Medical Center, Los Angeles, California 90048, United States; Advanced Clinical Biosystems Research Institute, Cedars-Sinai Medical Center, Los Angeles, California 90048, United States*

Aleksandra Binek − *Smidt Heart Institute, Cedars-Sinai Medical Center, Los Angeles, California 90048, United States; Advanced Clinical Biosystems Research Institute, Cedars-Sinai Medical Center, Los Angeles, California 90048, United States*

Zachary Dwight − *Precision Biomarker Laboratories, Cedars-Sinai Medical Center, Los Angeles, California 90048, United States*

Nethika R. Ariyasinghe − *Smidt Heart Institute, Cedars-Sinai Medical Center, Los Angeles, California 90048, United States; Advanced Clinical Biosystems Research Institute, Cedars-Sinai Medical Center, Los Angeles, California 90048, United States*

Sean Escopete − *Smidt Heart Institute, Cedars-Sinai Medical Center, Los Angeles, California 90048, United States; Advanced Clinical Biosystems Research Institute, Cedars-Sinai Medical Center, Los Angeles, California 90048, United States*

Sandy Y. Joung − *Smidt Heart Institute, Cedars-Sinai Medical Center, Los Angeles, California 90048, United States*

Susan Cheng − *Smidt Heart Institute, Cedars-Sinai Medical Center, Los Angeles, California 90048, United States*

Sarah J. Parker − *Smidt Heart Institute, Cedars-Sinai Medical Center, Los Angeles, California 90048, United States; Advanced Clinical Biosystems Research Institute, Cedars-Sinai Medical Center, Los Angeles, California 90048, United States;* orcid.org/0000-0001-8911-3615

Justyna Fert-Bober − *Smidt Heart Institute, Cedars-Sinai Medical Center, Los Angeles, California 90048, United States; Advanced Clinical Biosystems Research Institute, Cedars-Sinai Medical Center, Los Angeles, California 90048, United States;* orcid.org/0000-0002-2824-5056

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.2c00671

Author Contributions

[+]S.J.P., J.F.-B., and J.E.V.E. have equal contributions.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Erika Glazer for her generous support for the Erika J. Glazer Endowed Chair in Women's Heart Health, and funds from the Barbra Streisand Women's Heart Center, and the Smidt Heart Institute at Cedars-Sinai Medical Center as well as grants from National Institute of Health 5R01HL155346-01A1, 2R01HL111362-05A1, and 1R01HL144509-01.

REFERENCES

(1) Macklin, A.; Khan, S.; Kislinger, T. Recent advances in mass spectrometry based clinical proteomics: applications to cancer research. *Clin Proteomics.* **2020**, *17*, 17.

(2) Rana, R.; Rathi, V.; Ganguly, N. K. A comprehensive overview of proteomics approach for COVID 19: new perspectives in target therapy strategies. *J. Proteins Proteom.* **2020**, *11*, 223−232.

(3) Alves Martins, B. A.; de Bulhões, G. F.; Cavalcanti, I. N.; Martins, M. M.; de Oliveira, P. G.; Martins, A. M. A. Biomarkers in Colorectal Cancer: The Role of Translational Proteomics Research. *Front Oncol.* **2019**, *9*, 1284.

(4) Jacobs, J. M.; Adkins, J. N.; Qian, W.-J.; Liu, T.; Shen, Y.; Camp, D. G., 2nd; et al. Utilizing human blood plasma for proteomic biomarker discovery. *J. Proteome Res.* **2005**, *4*, 1073−1085.

(5) Poulos, R. C.; Hains, P. G.; Shah, R.; Lucas, N.; Xavier, D.; Manda, S. S.; et al. Strategies to enable large-scale proteomics for reproducible research. *Nat. Commun.* **2020**, *11*, 3793.

(6) Overmyer, K. A.; Shishkova, E.; Miller, I. J.; Balnis, J.; Bernstein, M. N.; Peters-Clarke, T. M.; et al. Large-Scale Multi-omic Analysis of COVID-19 Severity. *Cell Syst.* **2021**, *12*, 23−40.

(7) Fu, Q.; Kowalski, M. P.; Mastali, M.; Parker, S. J.; Sobhani, K.; van den Broek, I.; et al. Highly Reproducible Automated Proteomics Sample Preparation Workflow for Quantitative Mass Spectrometry. *J. Proteome Res.* **2018**, *17*, 420−428.

(8) Mc Ardle, A.; Binek, A.; Moradian, A.; Chazarin Orgel, B.; Rivas, A.; Washington, K. E.; et al. Standardized Workflow for Precise Mid- and High-Throughput Proteomics of Blood Biofluids. *Clin Chem.* **2022**, *68*, 450−460.

(9) Leek, J. T.; Scharpf, R. B.; Bravo, H. C.; Simcha, D.; Langmead, B.; Johnson, W. E.; et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **2010**, *11*, 733−739.

(10) Čuklina, J.; Lee, C. H.; Williams, E. G.; Sajic, T.; Collins, B. C.; Rodríguez Martínez, M.; et al. Diagnostics and correction of batch effects in large-scale proteomic studies: a tutorial. *Mol. Syst. Biol.* **2021**, *17*, No. e10240.

(11) Johnson, W. E.; Li, C.; Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics.* **2007**, *8*, 118−127.

(12) Wehrens, R.; Hageman, J. A.; van Eeuwijk, F.; Kooke, R.; Flood, P. J.; Wijnker, E.; et al. Improved batch correction in untargeted MS-based metabolomics. *Metabolomics.* **2016**, *12*, 88.

(13) Allison, D. B.; Cui, X.; Page, G. P.; Sabripour, M. Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.* **2006**, *7*, 55−65.

(14) Mecham, B. H.; Nelson, P. S.; Storey, J. D. Supervised normalization of microarrays. *Bioinformatics.* **2010**, *26*, 1308−1315.

(15) Evans, C.; Hardin, J.; Stoebel, D. M. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief Bioinform.* **2018**, *19*, 776−792.

(16) Benito, M.; Parker, J.; Du, Q.; Wu, J.; Xiang, D.; Perou, C. M.; et al. Adjustment of systematic microarray data biases. *Bioinformatics.* **2004**, *20*, 105−114.

https://doi.org/10.1021/acs.jproteome.2c00671
J. Proteome Res. XXXX, XXX, XXX−XXX

(17) Dubois, E.; Galindo, A. N.; Dayon, L.; Cominetti, O. Assessing normalization methods in mass spectrometry-based proteome profiling of clinical samples. *Biosystems.* **2022**, *215*−*216*, 104661.
(18) Välikangas, T.; Suomi, T.; Elo, L. L. A systematic evaluation of normalization methods in quantitative label-free proteomics. *Brief Bioinform.* **2018**, *19*, 1−11.
(19) Zhang, Y.; Parmigiani, G.; Johnson, W. E. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom Bioinform.* **2020**, *2*, lqaa078.
(20) Zhu, T.; Sun, R.; Zhang, F.; Chen, G.-B.; Yi, X.; Ruan, G.; et al. BatchServer: A Web Server for Batch Effect Evaluation, Visualization, and Correction. *J. Proteome Res.* **2021**, *20*, 1079−1086.
(21) Rudnick, P. A.; Markey, S. P.; Roth, J.; Mirokhin, Y.; Yan, X.; Tchekhovskoi, D. V.; et al. A Description of the Clinical Proteomic Tumor Analysis Consortium (CPTAC) Common Data Analysis Pipeline. *J. Proteome Res.* **2016**, *15*, 1023−1032.
(22) Shen, M.; Chang, Y.-T.; Wu, C.-T.; Parker, S. J.; Saylor, G.; Wang, Y.; et al. Comparative assessment and novel strategy on methods for imputing proteomics data. *Sci. Rep.* **2022**, *12*, 1067.
(23) Dabke, K.; Kreimer, S.; Jones, M. R.; Parker, S. J. A Simple Optimization Workflow to Enable Precise and Accurate Imputation of Missing Values in Proteomic Data Sets. *J. Proteome Res.* **2021**, *20*, 3214−3229.
(24) Röst, H. L.; Rosenberger, G.; Navarro, P.; Gillet, L.; Miladinović, S. M.; Schubert, O. T.; et al. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.* **2014**, *32*, 219−223.
(25) Demichev, V.; Messner, C. B.; Vernardis, S. I.; Lilley, K. S.; Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods.* **2020**, *17*, 41−44.
(26) Clough, T.; Thaminy, S.; Ragg, S.; Aebersold, R.; Vitek, O. Statistical protein quantification and significance analysis in label-free LC-MS experiments with complex designs. *BMC Bioinformatics.* **2012**, *13*, S6.
(27) Teo, G.; Kim, S.; Tsou, C. C.; Collins, B.; Gingras, A. C.; Nesvizhskii, A. I.; et al. mapDIA: Preprocessing and statistical analysis of quantitative proteomics data from data independent acquisition mass spectrometry. *J. Proteomics* **2015**, *129*, 108−120.
(28) Wei, R.; Wang, J.; Su, M.; Jia, E.; Chen, S.; Chen, T.; et al. Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. *Sci. Rep.* **2018**, *8*, 1−10.
(29) Ebinger, J. E.; Botwin, G. J.; Albert, C. M.; Alotaibi, M.; Arditi, M.; Berg, A. H.; et al. Seroprevalence of antibodies to SARS-CoV-2 in healthcare workers: a cross-sectional study. *BMJ. Open.* **2021**, *11*, No. e043584.
(30) Fu, Q.; Johnson, C. W.; Wijayawardena, B. K.; Kowalski, M. P.; Kheradmand, M.; Van Eyk, J. E. A Plasma Sample Preparation for Mass Spectrometry using an Automated Workstation. *J. Vis Exp.* **2020**, DOI: 10.3791/59842.
(31) Reiter, L.; Rinner, O.; Picotti, P.; Hüttenhain, R.; Beck, M.; Brusniak, M.-Y.; et al. mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nat. Methods.* **2011**, *8*, 430−435.
(32) Röst, H. L.; Liu, Y.; D'Agostino, G.; Zanella, M.; Navarro, P.; Rosenberger, G.; et al. TRIC: an automated alignment strategy for reproducible protein quantification in targeted proteomics. *Nat. Methods.* **2016**, *13*, 777−783.
(33) Cleveland, W. S.; Grosse, E.; Shyu, W. M. Local Regression Models. *Statistical Models in S.* **2017**, 309−376.
(34) Wright, M. N.; Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J. Stat. Softw.* **2017**, *77*, 1−17.
(35) Jin, Z.; Kang, J.; Yu, T. Missing value imputation for LC-MS metabolomics data by incorporating metabolic network and adduct ion relations. *Bioinformatics.* **2018**, *34*, 1555−1561.
(36) Choi, M.; Chang, C. Y.; Clough, T.; Broudy, D.; Killeen, T.; MacLean, B.; et al. MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics* **2014**, *30* (17), 2524−2526.
(37) Pang, Z.; Chong, J.; Zhou, G.; de Lima Morais, D. A.; Chang, L.; Barrette, M.; et al. MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights. *Nucleic Acids Res.* **2021**, *49*, W388−W396.
(38) Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; et al. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **2000**, *25*, 25−29.
(39) The Gene Ontology ConsortiumThe Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **2018**, *47* (D1), D330−D338.
(40) Du, J.; Yuan, Z.; Ma, Z.; Song, J.; Xie, X.; Chen, Y. KEGG-PATH: Kyoto encyclopedia of genes and genomes-based pathway analysis using a path analysis model. *Mol. Biosyst.* **2014**, *10*, 2441−2447.
(41) Fabregat, A.; Sidiropoulos, K.; Viteri, G.; Forner, O.; Marin-Garcia, P.; Arnau, V.; et al. Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinformatics* **2017**, *18*, 142.
(42) Slenter, D. N.; Kutmon, M.; Hanspers, K.; Riutta, A.; Windsor, J.; Nunes, N.; et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.* **2018**, *46*, D661−D667.
(43) Escher, C.; Reiter, L.; MacLean, B.; Ossola, R.; Herzog, F.; Chilton, J.; MacCoss, M. J.; Rinner, O. Using iRT, a normalized retention time for more targeted measurement of peptides. *Proteomics.* **2012**, *12* (8), 1111−21.
(44) Jin, L.; Bi, Y.; Hu, C.; Qu, J.; Shen, S.; Wang, X.; et al. A comparative study of evaluating missing value imputation methods in label-free proteomics. *Sci. Rep.* **2021**, *11*, 1−11.
(45) Wang, S.; Li, W.; Hu, L.; Cheng, J.; Yang, H.; Liu, Y. NAguideR: performing and prioritizing missing value imputations for consistent bottom-up proteomic analyses. *Nucleic Acids Res.* **2020**, *48*, No. e83.
(46) Schoof, E. M.; Furtwängler, B.; Üresin, N.; Rapin, N.; Savickas, S.; Gentil, C.; et al. Quantitative single-cell proteomics as a tool to characterize cellular hierarchies. *Nat. Commun.* **2021**, *12*, 1−15.

https://doi.org/10.1021/acs.jproteome.2c00671
J. Proteome Res. XXXX, XXX, XXX−XXX