Yu *et al. Genome Biology* (2024) 25:254
https://doi.org/10.1186/s13059-024-03401-9

Genome Biology

REVIEW

Open Access

Assessing and mitigating batch effects in large-scale omics studies

Ying Yu[1*], Yuanbang Mai[1], Yuanting Zheng[1*] and Leming Shi[1,2,3*]

*Correspondence:
ying_yu@fudan.edu.cn;
zhengyuanting@fudan.edu.cn;
lemingshi@fudan.edu.cn

[1] State Key Laboratory of Genetic Engineering, School of Life Sciences and Human Phenome Institute, Fudan University, Shanghai, China
[2] Cancer Institute, Shanghai Cancer Center, Fudan University, Shanghai, China
[3] International Human Phenome Institutes (Shanghai), Shanghai, China

Abstract

Batch effects in omics data are notoriously common technical variations unrelated to study objectives, and may result in misleading outcomes if uncorrected, or hinder biomedical discovery if over-corrected. Assessing and mitigating batch effects is crucial for ensuring the reliability and reproducibility of omics data and minimizing the impact of technical variations on biological interpretation. In this review, we highlight the profound negative impact of batch effects and the urgent need to address this challenging problem in large-scale omics studies. We summarize potential sources of batch effects, current progress in evaluating and correcting them, and consortium efforts aiming to tackle them.

Introduction

Batch effects are technical variations that are irrelevant to study factors of interest. They are introduced into high-throughput data due to variations in experimental conditions over time, using data from different labs or machines, or using data from different analysis pipelines [1–4]. Batch effects can be commonly seen in omics data, such as genomics [5–8], transcriptomics [4, 9, 10], proteomics, metabolomics [11], and multiomics integration [12, 13]. Recent advances in single-cell sequencing technology (i.e., scRNA-seq) have provided opportunities for resolving gene expression heterogeneity in single cells. Nevertheless, it brings more complex batch effects to arise [14, 15]. Batch effects can introduce noise that can dilute biological signals, reduce statistical power, or even result in misleading, biased, or non-reproducible results [3]. What is worse, batch effects can act as a paramount factor contributing to irreproducibility, resulting in retracted articles, invalidated research findings, and economic losses [16].

Batch effects are more complex in omics data because they involve multiple types of data that are measured on different platforms and have different distributions and scales [12, 17]. Multiomics profiling is a powerful tool for identifying differential features between biological groups based on multiple omics types [18, 19], and has demonstrated incredible potential in biomedical research to discover biomarkers

for clinical diagnosis, prognosis, and therapeutic action [20–24]. The rapid advancement of technology and the reduction in costs have made the analysis of multiomics data common in research. However, this has also led to an increase in the occurrence of batch effects [1]. With more researchers performing multiomics analyses, tackling batch effects in multiomic integration is urgently needed.

Furthermore, the challenges of batch effects are magnified in longitudinal and/or multi-center studies. Many longitudinal studies aim to determine how a time-varying exposure affects the outcome variable(s). However, technical variables may affect the outcome in the same way as the exposure. For example, sample processing time in generating omics data is probably confounded with the exposure time. Such scenarios are particularly problematic to identify features that change over time, because it is difficult or almost impossible to distinguish whether the detected changes are driven by time/exposure or caused by an artifact from batch effects [25].

Recently, single-cell technologies such as scRNA-seq have provided opportunities to gain in-depth insights into samples with heterogeneity. However, compared to traditional RNA-seq technologies, or named bulk RNA-seq, scRNA-seq suffers higher technical variations [26]. Specifically, scRNA-seq methods have lower RNA input, higher dropout rates, and a higher proportion of zero counts, low-abundance transcripts, and cell-to-cell variations than bulk RNA-seq [27]. These factors make batch effects more severe in single-cell data than in bulk data. Batch effects and the selection of correction algorithms have been shown to be predominant factors in large-scale and/or multi-batch scRNA-seq data [14, 15, 26].

Despite extensive research and discussions on developing and comparing batch effect correction algorithms (BECAs), finding solutions for tackling batch effects is still an active research topic. One possible reason is that the disparity in nature of the batch effects makes it difficult to have a one-fit-all tool. New BECAs continue to be developed, presenting a bewildering choice of various BECAs to investigators. Investigators may get confused in choosing a proper method, each with its own set of capabilities and limitations. One could argue that the underlying cause of batch effects might have not yet been correctly identified, leading to conflicting or confusing conclusions in this field.

There have also been extensive reviews written on issues of batch effects, in RNA-seq/microarray [4], scRNA-seq [1], proteomics [2, 28], metabolomics [11], and multiomics [3, 29]. A systematic review of the topic at the omics level is still much needed, due to the complexity of batch effects across omics types and what could be learned from the commonalities of batch effects across omics types. Previous research has shown that some issues are shared across various omics types, while others are specific to certain fields [17]. Consequently, several BECAs that were originally developed based on one omics type are shown to be applicable to other omics types [30, 31], while others are applicable to certain omics type(s) as they were developed to address platform-specific problems [32]. With the rapid advancement of technology, the field of batch effects is rapidly evolving. Although there is already a significant body of research on this topic of batch effects, the idea of batch effects in omics data is not adequately addressed. Therefore, the field needs more work to handle the complexity and diversity of large-scale, multiomics data. A comprehensive review of the topic at the omics level can help investigators better

understand the potential sources of batch effects, and implement appropriate strategies to minimize or correct them.

In this review, we first highlight the profound negative impact of batch effects and the continuous need to address this problem. Next, we review and discuss potential sources of batch effects, the current progress of diagnostics and correction of batch effects, and consortium efforts to harness batch effects. Finally, we discuss current challenges and future directions to tackle the batch effect problem and push forward multiomics integration.

Profound negative impact of batch effects

Batch effects may lead to incorrect conclusions

Batch effects have profound negative impacts. In the most benign cases, batch effects will lead to increased variability and decreased power to detect a real biological signal. Batch effects can also interfere with downstream statistical analysis. Batch-correlated features can be erroneously identified in differential gene expression analysis [33–35] and prediction [36], especially when batch and biological outcomes are highly correlated. In some worse cases, batch effects are correlated with one or more outcomes of interest in an experiment, affecting the interpretation of the data and leading to incorrect conclusions.

One example is that, in a clinical trial study, batch effects were introduced by a change in the RNA-extraction solution that was used in generating gene expression profiles, resulting in a shift in the gene-based risk calculation. This further resulted in incorrect classification outcomes for 162 patients, 28 of whom received incorrect or unnecessary chemotherapy regimens [37].

In another example, the cross-species differences between human and mouse were reported to be greater than the cross-tissue differences within the same species [38]. However, a more rigorous analysis of the data showed that data of human and mouse came from different subject designs, and the data generation timepoints were different by 3 years [39]. Batch effects were responsible for the so-called differences between human and mouse species. After batch correction, the gene expression data from human and mouse tended to cluster by tissue rather than by species [39].

Batch effect is a paramount factor contributing to irreproducibility

Reproducibility is a fundamental requirement in scientific research and there has been a growing concern among both scientists and the public on the lack of reproducibility [40–42]. A survey conducted by Nature found that 90% of respondents (1576 respondents surveyed) believed that there was a reproducibility crisis, with over half considering it a significant crisis [42]. Among the massive factors contributing to irreproducibility, batch effect(s) from reagent variability and experimental bias are paramount factors [16, 42].

Irreproducibility caused by batch effects can also result in rejected papers, discredited research findings, and financial losses [43]. Many high-profile articles were retracted due to batch-effect-driven irreproducibility of the key results [44, 45]. For instance, the authors of a study published in Nature Method identified a genetically encoded, fluorescent serotonin biosensor with high affinity and specificity [46]. However, the authors

later noticed that the sensitivity of the biosensor was highly dependent on the reagent batch, especially the batch of fetal bovine serum (FBS). When the batch of FBS was changed, the key results of the article could not be reproduced. The article was therefore retracted [44]. Moreover, despite overcoming many barriers and challenges [47], RPCB (Reproducibility Project: Cancer Biology) team still failed to reproduce over half of high-profile cancer studies [40, 41], highlighting the importance of eliminating batch effects across laboratories and making scientific results reproducible.

Sources and possible solutions for addressing batch effects

The fundamental cause of batch effects can be partially attributed to the basic assumptions of data representation in omics data [13]. In biomedical research, the concentration or abundance (C) of an analyte in a sample is crucial, and measurement technologies aim to provide the information. In quantitative omics profiling, the absolute instrument readout or intensity (I)—such as FPKM, FOT, or peak area, regardless of any per-sample normalization method applied—is often used as a surrogate for C. This relies on the assumption that under any experimental conditions, there is a linear and fixed relationship (f, or sensitivity) between I and C, expressed as $I = f(C)$. However, in practice, due to differences in diverse experimental factors, the relationship f may fluctuate. These fluctuations make I inherently inconsistent across different batches, leading to inevitable batch effects in omics data [13].

The occurrence of batch effects can be traced back to diverse origins and can emerge at every step of a high-throughput study. Despite the fact that some sources are common to numerous omics types, some are exclusive to particular fields. In-depth discussions have been conducted to elaborate on the sources of batch effects in genomics [48], transcriptomics [35, 49], proteomics [28], and metabolomics [11]. Here we highlight some of the most encountered sources of cross-batch variations during different phases of a typical high-throughput study (Fig. 1 and Table 1).

Study design

During the study design stage, some sources of batch effects can be introduced, including the choice of high-throughput technology, sample size, and number of batches. Among them, it has been reported that flawed or confounded study design is one of the critical sources of cross-study irreproducibility [42, 50]. This can happen if the samples are not collected in a randomized manner or if they are selected based on a specific characteristic, such as age, gender, or clinical outcome. This can lead to systematic differences between the batches, which can be difficult to correct for during data analysis. Another factor that is related to batch effects is the degree of treatment effect of interest [51]. If the degree of treatment effect of interest is minor, the expression profiles would be more susceptible to technical variations.

Sample preparation and storage

Variables in sample collection, preparation, and storage may introduce technical variations and affect the results of high-throughput profiling. These variables include protocol procedures, reagent lots, storage conditions, operators, and labs (Fig. 1).

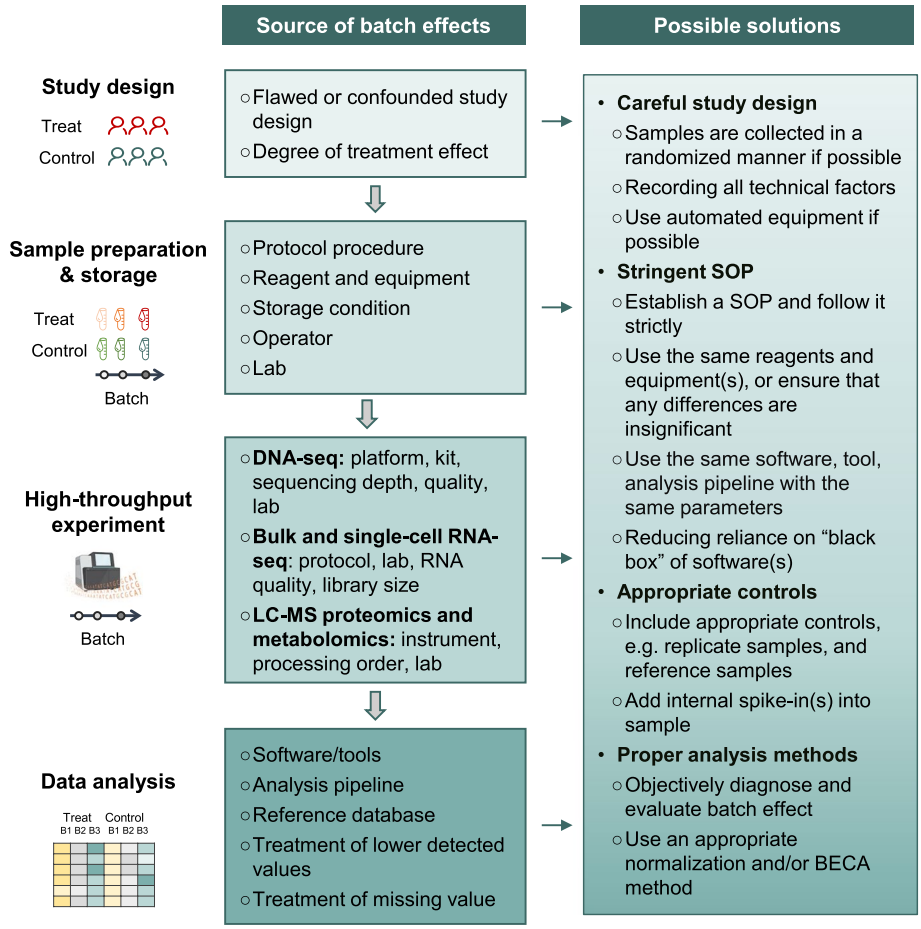

Fig. 1 Sources and possible solutions for addressing batch effects

Protocol procedure is one of the most important sources of cross-batch variations [52]. For example, plasma is widely used in biomarker discovery due to its easy accessibility [53, 54]. However, the biospecimens are likely exposed for varying periods of time and temperatures prior to centrifugation for plasma separation, which may cause significant changes in proteins [55] and metabolites [56–58]. Moreover, different blood processing protocols, such as different centrifugal forces, may result in different quantifications of plasma mRNA [59, 60]. These situations are likely to occur in large-scale studies when samples are collected at multiple centers/biobanks to fulfill the sample-size requirement.

Other factors, such as reagents, equipment, sample storage conditions, operators, and labs can also lead to variability in the quality and quantity of the samples across batches. For example, tubes that are coated with anticoagulants are widely used for the storage of blood samples. However, it has been reported that different types or concentrations of anticoagulants can result in differences in proteomics and metabolomics profiling of blood samples [11, 55]. Storage conditions, e.g., the numbers of freeze/thaw cycles, are sometimes overlooked, but samples stored under different conditions may have systematic differences in molecular profiling [61]. Moreover, the operator is another important contributor. When samples are processed by manual pipetting, there is a risk of personnel variability due to differences in pipetting techniques [62–64]. When different batches

Table 1 Potential sources of batch effects for different omics data types

Sources	Stage	Common or specific omics type	Description
Flawed or confounded study design	Study design	Common	This can happen if the samples are not collected in a randomized manner or if they are selected based on a specific characteristic, such as age, gender, or clinical outcome.
Degree of treatment effect of interest	Study design	Common	A minor treatment effect size makes it more difficult to distinguish from batch effects compared to large treatment effects.
Protocol procedure	Sample preparation and storage	Common	Different centrifugal forces during plasma separation, or time and temperatures prior to centrifugation, may cause significant changes in mRNA, proteins, and metabolites.
Sample storage conditions	Sample preparation and storage	Common	Variations in sample storage temperature, duration, freeze-thaw cycles, *etc.* can impact stability and introduce batch effects.
RNA enrichment protocol	High-throughput experiments	Bulk and single-cell RNA-seq	Distinct transcriptomes are represented in libraries prepared by different protocols, in particular expression profiles of non-polyadenylated transcripts, 3'UTRs and introns.
Instrument variability	High-throughput experiments	LC-MS proteomics and metabolomics	Instrument variability can arise from differences in the performance of the mass spectrometers, chromatography systems, and other instrumental factors. It leads to increased variability and decreased power to detect biologically meaningful responses.
Signal drift	High-throughput experiments	LC-MS proteomics and metabolomics	Signal drift is the gradual change in the intensity of the detected signal over time.
Analysis pipeline	Data analysis	Common	Use of different algorithms or parameters in the computational pipeline induces batch effects.
Reference database	Data analysis	Common	The alignment to different reference databases may lead to different results, because reference databases vary greatly in terms of their curation, completeness and comprehensiveness

Table 1 (continued)

Sources	Stage	Common or specific omics type	Description
Treatment of missing values	Data analysis	LC-MS proteomics and metabolomics	The methods of treatment of missing values, e.g., removing all features with missing values, filling with zeros or randomly small values or re-quantification/prediction based on different algorithms, can introduce bias and aggravate batch variations.
Reagents	Multiple	Common	Variations in the quality, composition, or performance of reagents used during sample preparation, processing, and data acquisition can introduce systematic technical biases, leading to batch-specific patterns. Differences in reagent lots, manufacturers, or expiration dates can further exacerbate these issues, making careful selection and consistent usage of reagents across all samples essential to minimize batch effects.
Equipment	Multiple	Common	Different equipment used for sample processing or data acquisition, or variations in the performance and calibration of the same equipment over time, can introduce systematic biases that manifest as batch effects.
Lab	Multiple	Common	Environmental factors such as temperature, humidity, and air quality within the laboratory can vary over time and across different lab spaces, potentially impacting sample stability, enzymatic reactions, and instrument performance.
Operators	Multiple	Common	When samples are processed by manual pipetting, there is a risk of personnel variability due to differences in pipetting techniques.

of samples are processed by different operators independently, the cumulative differences in sample volumes may become nonnegligible.

High-throughput experiments

DNA sequencing

One of the main sources of batch effects in DNA sequencing is the use of different sequencing platforms, which can lead to differences in the quality and quantity of the

sequencing data between batches. There is a vibrant and diverse market for sequencing platforms, including Illumina HiSeq and NovaSeq, ThermoFisher Ion Torrent, BGISEQ-500 and MGISEQ-2000, the GenapSys GS111, Oxford Nanopore Technologies (ONT) Flongle, MinION and PromethION flow cells, and PacBio CCS platforms, to name a few, with the diversity of cost, throughput, speed, sequence lengths, error rates, and bias [48, 65]. Other sources, such as experiment kits (i.e., exome capture kits for whole exome sequencing), sequencing depth, sequencing quality, and sequencing labs were showed variations between batches [5–7, 66].

Bulk and single-cell RNA-seq

For bulk and single-cell RNA-seq, one of the major sources is protocol. Specifically, for bulk RNA-seq, poly-A enrichment and ribosomal RNA (rRNA) depletion are two common protocols used to enrich mRNA from total RNA samples in RNA-seq experiments. Poly-A enrichment protocol involves capturing the poly-A tail of mRNA molecules using oligo(dT) beads, while rRNA depletion protocol involves removing rRNA molecules from the total RNA samples. Differences in RNA enrichment protocols can result in differences in the RNA population that is captured [67, 68]. Distinct transcriptomes are represented in libraries prepared by different protocols, in particular expression profiles of non-polyadenylated transcripts, 3'UTRs and introns [35], which can contribute to batch effects [69]. Moreover, diverse RNA extraction and library construction protocols are used in scRNA-seq, resulting in highly sensitive technical variability and biological heterogeneity, which can lead to batch effects [26]. Variabilities in any step of the scRNA-seq protocol can introduce batch effects. For example, differences in the efficiency of reverse transcription, the amount of cDNA amplified, or the quality of sequencing reads can lead to batch effects in downstream analyses [14]. Several publications have compared and reviewed scRNA-seq protocols in detail [70–72].

Lab is another important contributor to batch effects in RNA-seq. Investigators from the Sequencing Quality Control (SEQC) consortium examined three sequencing platforms at multiple laboratory sites using reference RNA samples with built-in controls and observed differences across labs and platforms [34, 73]. Recently, we performed a multi-lab RNA-seq experiment based on Quartet RNA reference materials, a suite of four RNA samples derived from immortalized B-lymphoblastoid cell lines from a family quartet of parents and monozygotic twin daughters, and found a vast diversity of expression profiles across labs in both poly-A and RiboZero protocols [69]. Similarly, strong lab effects were reported in scRNA-seq data using reference samples [14].

Additionally, other sources of batch effects in RNA-seq and scRNA-seq involve RNA quality [74], RNA purity [75], library size [49], sequencing platforms [35], etc.

LC-MS proteomics and metabolomics

Instrument variability is a major source of batch effects in LC-MS proteomics and metabolomics experiments [76, 77]. This can arise from differences in the performance of the mass spectrometers, chromatography systems, and other instrumental factors.

Signal drift within the instrument makes sample processing order another factor of batch effects of LC-MS technology [2, 11, 78]. Signal drift is the gradual change in the intensity of the detected signal over time due to various factors such as fluctuations in

LC performance, variations in electrospray process, changes in ion transfer caused by fouled or moved optics, and changes in detector sensitivity [79]. When samples are processed in batches, the processing order can influence the degree of signal drift, leading to batch effects in the data.

Moreover, lab differences were also observed in proteomics [80–82] and metabolomics experiments [83, 84], resulting from large variations of precursor mass-to-charge ratio (m/z) of the ion and retention times (rt) across labs [85].

Data analysis

Throughout the entire analytical workflow, data analysis has the inherent potential to introduce technical or unwanted variations at each juncture. Different analysis methods clearly lead to increased variability, and results may be different due to the analysis approach. It should be noted that the analytic variability can be avoided when applying the same analysis method to all the data. However, with the widespread adoption of high-throughput sequencing, especially in multi-center, long-term longitudinal studies or clinical applications, it is not rare that different processing methods are applied to different batches of data, especially when the raw data are not available. In these cases, different analysis methods can become a potential contributor to variations and broadly regarded as a source of batch effects, as have been reported across diverse fields, such as genomics [66, 86, 87], transcriptomics [73, 88], proteomics [89], and metabolomics [11].

For genomics data, Pan et al., conducted whole genome sequencing (WGS) of the same eight DNA samples from three library kits in six labs and called variants with 56 combinations of aligners and callers [87]. Bioinformatics pipelines (callers and aligners), together with sequencing platform and library preparation influenced the germline mutation detection. Among them, bioinformatics pipelines have shown a larger impact. Similar results were also observed by O'Rawe et al., in terms of whole-exome and genome sequencing data [90].

For bulk RNA-seq data, the SEQC consortium found that data analysis pipelines, including gene quantification, junction identification, and differential expression contributed to measurement performances and variations were large for transcript-level profiling [73]. Similarly, Sahraeian et al. constructed a comparative study for RNA-seq workflows, by assessing 39 analysis tools with ~120 combinations and finding a diversity of performances in terms of read alignment, assembly, isoform detection, quantification, RNA editing, and RNA-seq-based variant calling [88].

For LC-MS proteomics and metabolomics data, the use of different searching methods to decode tandem mass spectra and match them to databases of theoretical tryptic peptides or metabolites is a source of variability, because of differences in the searching tools with different false discovery rates. Furthermore, the peak alignment to different reference databases may lead to different results, because reference databases vary greatly in terms of their curation, completeness, and comprehensiveness [81]. In addition, because missing values are common in proteomics and metabolomics data and are batch- and feature-specific, the methods of treatment of missing values, e.g., removing all features with missing values, filling with zeros or randomly small values or re-quantification/prediction based on different algorithms, can introduce bias and aggravate batch variations [2, 80].

Possible solutions

Some possible solutions can be applied to minimize batch effects in the high-throughput study (Fig. 1). Notably, some of these solutions are in line with the principle of reproducibility in scientific research [16, 50, 91].

First, careful study design will somehow be effective. It is important to ensure that the samples are collected in a randomized manner if possible. If the degree of treatment effect of interest is minor in the study, the study may be more susceptible to batch effects, and researchers should pay more attention to study design, such as setting more replicates and choosing more rigorous measurement technologies. It should be noted that randomization in study design is ideal but almost impossible in reality. In longitudinal and multi-center studies, it is inevitable for a study design to include confounded batches and biological factors. On the other hand, even in a perfectly designed study, batches will still be introduced, because the experiments may span a long period of time or involve personnel changes. In these cases, recording as many technical factors as possible can be useful in the following analysis, including diagnosis and correction of batch effects, as described below. In addition, the use of automated sample preparation systems can also help minimize variability between batches.

Secondly, standard operating procedures (SOPs) shall be established and validated at the beginning of a large-scale study, with strict adherence by all operators and technicians. This also includes using the same wet lab conditions (reagents, equipment, etc.), as well as the same dry lab conditions (analysis pipeline with the same parameters, software, etc.). If achieving these is challenging, it is crucial to conduct an objective assessment to ensure that any differences between these variations are insignificant.

Thirdly, appropriate controls should be included to help evaluate and correct batch effects, and further improve intra- and inter- batch reproducibility. For example, investigators can involve replicate samples or reference samples within each batch. Adding spike-ins to study samples can also be applied. Multicomics reference samples and spike-in products have been reviewed in our accompanying work [92].

Finally, proper analysis methods should be applied to mitigate the effects, including diagnostics and evaluation methods, normalization, and/or BECA method(s), which have been mentioned in the next sections.

Diagnostics and evaluation of batch effects

Prior to performing BECAs, diagnostics and evaluation is needed to understand the existence of batch effects and the estimation of the proportion of variation in the data resulting from batch effects. Additionally, evaluation is also needed to be applied after performing BECAs to estimate whether batch effects have been successfully removed. These evaluation steps are particularly important, because some BECAs should only be conducted when necessary, and serious errors might be introduced when improper BECAs are used [3].

The evaluation of batch effects can be performed not only based on expression profiles, but also based on quality control metrics [93], such as read coverage [29, 48], GC content [48], nucleotide composition [8], mapping rate [48], and mismatch rate [48]. Batch effects in quality control metrics may further affect data processes such as data

filtering, normalization, and interpretation. Some tools, for example, BatchQC, were developed for facilitating diagnostics and evaluation of batch effects [94].

In this section, we first describe a typical workflow regarding when and why evaluation methods can be performed. We subsequently provide a dozen of visualizations and measurements to specifically show how to perform the evaluation methods. It should be noted that, most methods for evaluating batch effects are focused on quantity omics rather than qualitative omics such as WGS. Methods for determining whether batch effects exist in qualitative omics data warrant further investigations.

Typical workflow

In a typical analysis workflow, five major stages can be applied to transform the raw data matrix into a finalized data matrix that is ready for downstream analyses. This includes initial assessments, normalization, diagnostics of batch effects, batch correction, and assessment of the correction of batch effects (Fig. 2a).

Initial assessments can be applied based on the raw data matrix to determine the size of biases. Specifically, data quality assessment can be applied to check for any errors, inconsistencies, missing values, outliers, or noise in the data and correct them if possible. Additionally, data exploration tools can be applied to summarize the main characteristics of the data, such as the number of variables, the range of values, the distribution of values, and the correlation between variables.

According to the data quality and structure learned from the initial assessment, a normalization step can be performed to adjust distributional differences across samples and make samples more comparable in the global pattern. Meanwhile, normalization is to make the data to have a normal distribution or a unit norm to facilitate downstream statistical analyses, which is quite common in omics analysis. Dozens of normalization methods for correcting experimental variations and biases in high-throughput data have been developed, which have been extensively discussed in the reviews [95, 96].

Because batch effects may affect different genes in different ways [29] and normalization does not necessarily remove batch effects, diagnostics of batch effects can be applied to identify the source and patterns of batch effects and select an appropriate BECA. If batch effects exist in the datasets, batch correction can be applied to address feature-specific biases, which are summarized in the next section. It should be noted that normalization may eliminate the need for additional data correction. This can be confirmed through diagnostic plots and measurements, as described below. If the results after normalization are satisfactory, it is recommended to minimize data manipulation [2].

Finally, assessment of the correction of batch effects can be applied to test whether batch effects have been successfully migrated while retaining biological signals of interest. Sometimes it is not easy because true biological signals are probably previously unknown.

Visualization

One of the most common methods for diagnosing and evaluating batch effects is visualization, which provides an initial impression of the effectiveness of BECAs. To better illustrate visualization tools in terms of assessing batch effects, we employ a multi-batch

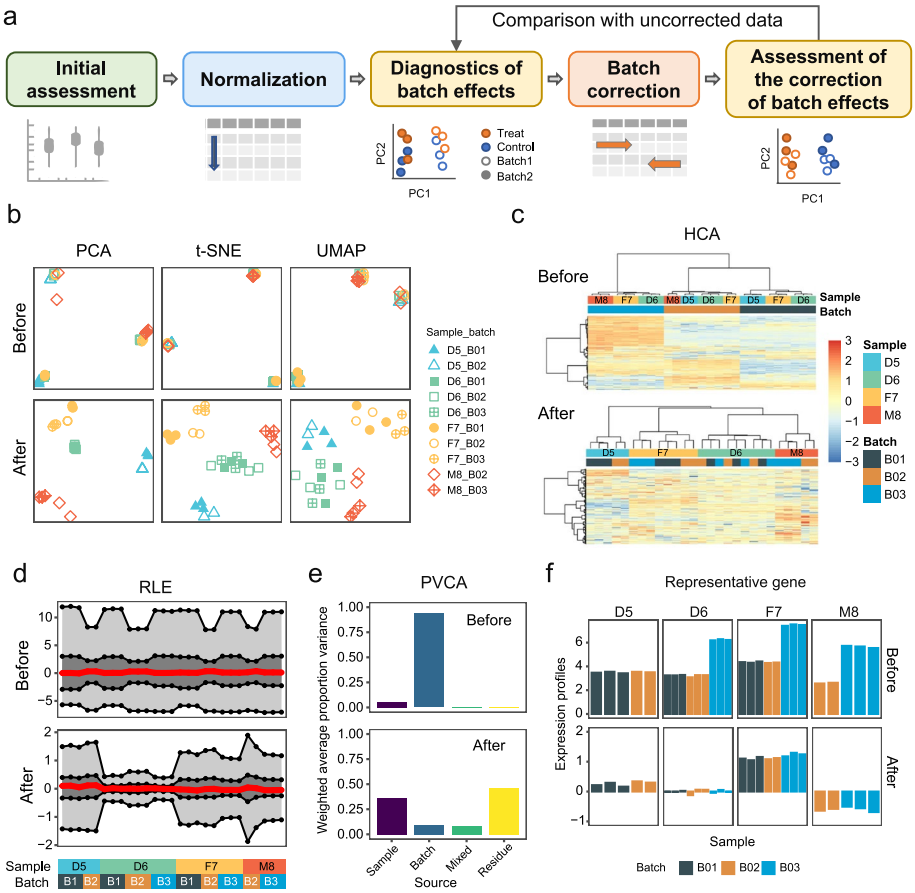

Fig. 2 Diagnostics and visualization of batch effects. **a** Overview of a typical analysis workflow. **b–f** Examples of visualization plots before and after batch correction. The dataset used for visualization is an RNA-seq dataset from four Quartet RNA reference materials [69], including 27 libraries from three batches. The examples have been performed using ratio-based scaling as the method for batch effect removal. **b** Dimensionality reduction methods, including principal component analysis (PCA), T-distributed stochastic neighbor embedding (t-SNE), and uniform manifold approximation and projection (UMAP). **c** Hierarchical clustering analysis (HCA). **d** Relative log expression (RLE) plot. **e** Principal variance component analysis (PVCA) coupled with bar plots. **f** The expression pattern of a representative gene across batches

RNA-seq dataset of four Quartet RNA reference materials [69], including 27 libraries from three batches. Different numbers of replicates (n=5~9) of reference materials are included in each batch to mimic a confounded scenario where replicates of reference materials are not equally distributed across batches. Data are available at Open Archive for Miscellaneous Data (OMIX) (accession number: OMIX002254) [97]. The examples have been performed using ratio-based scaling as the method for batch effect removal (Fig. 2b–f). Detailed information of the dataset, as well as code for reproducing the analysis has been deposited on GitHub [98].

First, dimensionality reduction methods are the most widely used visualization methods to identify the major sources of variation in high-dimensional data, including linear-based method(s) such as principal component analysis (PCA), and non-linear-based methods such as T-distributed stochastic neighbor embedding (t-SNE) [99] and uniform manifold approximation and projection (UMAP) [100, 101]. If batch effects exist,

samples will tend to be grouped by batches (Fig. 2b). It should be noted that t-SNE and UMAP are good at revealing local structures in high-dimensional data but cannot preserve the global structure of the data, which means that the relative distances and positions between clusters produced by the two methods are less meaningful [102].

Secondly, hierarchical clustering analysis (HCA) can be applied to show the clustering of the data by batches or by biological groups and indicate the presence or absence of batch effects (Fig. 2c). HCA is a dendrogram algorithm that groups similar samples into a cluster tree. 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.

Thirdly, relative log expression (RLE) plot which shows the distribution of the log-ratios of each gene's intensity over its geometric mean across all samples can help detect batch effects by comparing RLE values across different batches or groups of samples [103]. If there is a batch effect, the RLE plot may display the batch-specific distributions of medians or variances (Fig. 2d).

Fourthly, principal variance component analysis (PVCA) coupled with bar plots can be used for quantifying and visualization of the proportion of variations of experimental effects including batch (Fig. 2e). PVCA leverages the strengths of two methods to estimate the variance components: PCA and variance component analysis (VCA). PCA finds low-dimensional linear combinations of data with maintaining maximal variability, whereas VCA analysis attributes and partitions variability into known sources through a mixed linear model [104].

Finally, one straightforward way is to plot the expression patterns of individual features across batches. Technical factors (e.g., batch, processing order) can be used in the x-axis, and expression profiles can be used in the y-axis (Fig. 2f).

Measurements

While visualization alone may not provide a comprehensive evaluation, applying quantitative measurements is necessary to accurately assess the batch effect removal process. The following quantitative measurements may be employed for evaluating batch effects (Table 2).

First, distance-based metrics are proposed to calculate sample-wise distances to measure the similarity of samples across batches, such as alignment score [105], Distance ratio score (DRS) [106], Guided PCA [107], k-nearest neighbor batch-effect test (kBET) [108], Local inverse Simpson's index (LISI) [15, 109], Shannon Entropy [110], and signal-to-noise ratio (SNR) [13]. kBET is a widely used metric in scRNA-seq and is used to measure the batch mixing at the local level of the k-nearest neighbors [108]. kBET is easy to implement and sensitive to detect small batch effects. However, the disadvantages of kBET include its inability to work if class or batch proportions are highly confounded, if extreme outliers are present, or if high data-specific heterogeneity is present [1, 108]. Additionally, we previously proposed a metric called SNR for quantifying the ability to separate distinct biological groups when multiple batches of data were integrated [13, 32]. SNR is calculated based on PCA and measures the ability to differentiate intrinsic biological differences among distinct groups ("signal") from technical variations

Table 2 Representative measurements for evaluating batch effects

Name	Data type	Category	Description	Refs
Alignment score	scRNA-seq	Distance	Alignment score examines the local neighborhood of each cell after alignment. It is calculated using a nearest-neighbor graph based on the cells' embedding in some low-dimensional space.	[105]
Distance ratio score (DRS)	microarray	Distance	For a sample of a certain sample type, distance ratio score is the log of the ratio of the distance to the closest sample of a different sample type to the distance to the closest sample belonging to a different batch but the same sample type. The DRS is high if samples of the same type cluster together irrespective of batch since the denominator will be small compared to the numerator.	[106]
Guided PCA (gPCA)	RNA-seq, copy number alteration, methylation	Distance	gPCA is an extension of PCA to quantify and visualize the existence of batch effects. gPCA is guided by a batch indicator matrix using the singular-value decomposition (SVD) algorithm to look for batch effects in the data.	[107]
k-nearest neighbor batch-effect test (kBET)	scRNA-seq	Distance	kBET uses a Chi-based test for random neighborhoods of fixed size to determine whether they are well mixed, followed by an averaging of the binary test results to return an overall rejection rate. This result is easy to interpret: low rejection rates imply well-mixed replicates.	[108]
Local inverse Simpson's index (LISI)	scRNA-seq	Distance	LISI is a diversity index to measure the diversity of gene expression in scRNA-seq data and identify areas of high diversity that may be affected by batch effect. LISI first selects neighbors based on the local distribution of distances with a fixed perplexity. The selected neighbors are then used to compute the inverse Simpson's index for diversity, which is the effective number of types present in this neighborhood.	[15, 109]

Table 2 (continued)

Name	Data type	Category	Description	Refs
Shannon Entropy	scRNA-seq	Distance	The entropy-based metric is computed as follows: a k-NN graph is constructed based on the normalized data using Euclidean distance. The distribution of individuals in the neighborhood of each cell is then computed. Shannon entropy is further computed as a measure of diversity, resulting in one entropy value per cell.	[110]
Signal-to-noise ratio (SNR)	RNA-seq, proteomics, metabolomics, miRNA-seq, methylation	Distance	SNR is defined as the ratio of the average distance among different samples to the average distance among technical replicates of the same sample. Based on principal component analysis (PCA), distances between two samples in the space defined by the first two PCs were used to represent distances between the two samples. A higher SNR value indicates a lower technical effect in the data.	[13]
Adjusted rand index (ARI)	multiple	Cluster	ARI measures the similarity of the true labels and the clustering labels while ignoring permutations with chance normalization, which means random assignments will have an ARI score close to zero. ARI is in the range of -1 to 1, with 1 being the perfect clustering.	[111]
Average silhouette width (ASW)	scRNA-seq	Cluster	The calculation of a silhouette aims to determine whether a particular clustering has minimized within-cluster dissimilarity and maximized inter-cluster dissimilarity.	[112]
P_{low} from seqQscorer	RNA-seq, ChiP-seq, DNase-seq	Cluster	A metric P_{low} is a machine-learning derived probability for a sample to be of low quality, as derived by the seqQscorer tool. If P_{low} scores between the batches are significantly different, it means there are batch effects.	[74, 113]

Table 2 (continued)

Name	Data type	Category	Description	Refs
Performances of identifying DEFs	Multiple	Downstream	Comparison with the truly differentially or non-differentially expressed features, using metrics such as true positive rate (TPR), true negative rate (TNR), precision, recall, Matthews correlation coefficient (MCC), etc.	
Performance of predictive modeling	Multiple	Downstream	Classification models: ROC curve, confusion matrix Regression endpoint: mean squared error (MSE), mean absolute error (MAE), R-squared	

including batch effects of the same groups ("noise"). Generally, a higher SNR value indicates higher distinguishing power, and vice versa.

Secondly, cluster-based metrics are proposed to calculate the clustering accuracy or similarity against the batch effects, such as adjusted rand index (ARI) [111], average silhouette width (ASW) [112], and P_{low} from seqQscorer software [113]. ARI measures the similarity between the true labels and the clustering labels and reducing the influence of random permutations, which means random assignments will have an ARI score close to zero. The meaning of ARI depends on the setting of true class labels of the samples. A larger value of ARI with biological groups as the true groups value means better performance, while a smaller value of ARI batch denotes better batch effect correction.

Thirdly, as differential expression and prediction are two important downstream analysis tasks for quantitative omics, evaluations based on these tasks can demonstrate the need and effectiveness of BECAs with respect to biological interest, as performed in [32, 36, 114] and reviewed in [4]. Cross-batch results can be compared with the true set (i.e., truly differentially or non-differentially expressed features) to evaluate the performance of differential expression. Various metrics can be applied, including metrics based on the confusion matrix (i.e., true positives, true negatives, false positives, and false negatives), such as sensitivity or true positive rate (TPR), specificity or true negative rate (TNR), the positive predictive value (PPV), the negative predictive value (NPC), accuracy (ACC), and the Matthews correlation coefficient (MCC). Moreover, numerical metrics can also be applied, for example, the correlation coefficient representing the consistency of fold-changes with the true set, and the root mean square error (RMSE) representing the distance with the true set in fold-changes [69]. It should be noted that the determination of the true set is important before evaluations. The inclusion of a proper size of true set representing clinical purposes is preferred. The large size of the true set makes the evaluation easy and straightforward, but it does not help evaluate the ability to detect subtler differential expression for clinical purposes.

Moreover, cross-batch prediction is a critical aspect of multiomics analysis, particularly when it comes to identifying and validating molecular expression signatures that can be used for diagnosis, prognosis, and prediction of diseases and subsequent biomarker development [36]. In many cases, a predictive model is built using a batch of

samples (existing data), which is then applied to other batches of samples (future data). These datasets may be confounded with batch effects, which can negatively impact prediction by obscuring and washing out any predictive power of useful biological variations between certain outcomes [36, 114, 115]. Cross-batch prediction results can be compared with the truth (e.g., clinical endpoint) to evaluate the performance of prediction. There are various metrics that can be applied. For classification models, metrics based on the confusion matrix listed above and area under curve (AUC)) can be applied [115], while for regression models, mean squared error (MSE), mean absolute error (MAE), and R-squared can be used [116].

Additionally, Zhang et al. presented moment-based metrics for interrogating the shape of the distribution of batches to determine how batch effect should be adjusted [117].

Currently available batch-effect correction algorithms (BECAs)

When batch effects are confirmed to exist in datasets, actions are needed to be taken to avoid confounding effects in data analysis. Various strategies have been proposed to correct or minimize batch effects. Here, BECAs can be classified into four categories based on their underlying assumptions, including location-scale (LS) methods, matrix-factorization (MF) methods, distance-neighborhood based (DN) methods, and deep-learning (DL) methods (Fig. 3 and Additional file 1).

Location-scale (LS) methods

LS methods assume a statistical model for the location (mean) and/or scale (variance) of the data within the batches and proceed to adjust the batches in order to agree with these models [4].

ComBat is one of the most widely used BECAs in transcriptomics [118, 119] proteomics [120], and metabolomics [121]. It uses an empirical Bayes framework to estimate the magnitude (mean and variance) of batch effects, and then remove them [31]. Several extensions of the ComBat method have been developed. For example, modified ComBat (M-ComBat) transforms all feature distributions into a pre-determined "reference" batch, instead of the overall mean, providing more flexibility [122]. This reference-batch approach not only effectively corrects for batch effects but also holds the potential to facilitate validation of newly discovered biomarkers while enhancing predictions of pathway activities and drug effects [117]. A regularized version of ComBat (reComBat) [123] replaces a linear regression with a regularized linear regression model to handle highly correlated batch-sample situations. ComBat-seq, an extended version of ComBat for RNA-seq count data, retains the integer nature of count data and make the batch-adjusted data compatible with software packages that require integer counts [124].

Another example is the ratio-based method, i.e., by scaling the absolute feature values of study samples relative to those of concurrently profiled reference material(s). The ratio-based method is broadly effective in multiomics datasets, especially when batch effects are completely confounded with biological factors of study interests [32]. On the other hand, while the ratio-based method performed favorably in both balanced and confounded scenarios, it is not free of limitations. The ratio-based method is possible when the introduction of a reference sample can be decided as part of the experimental

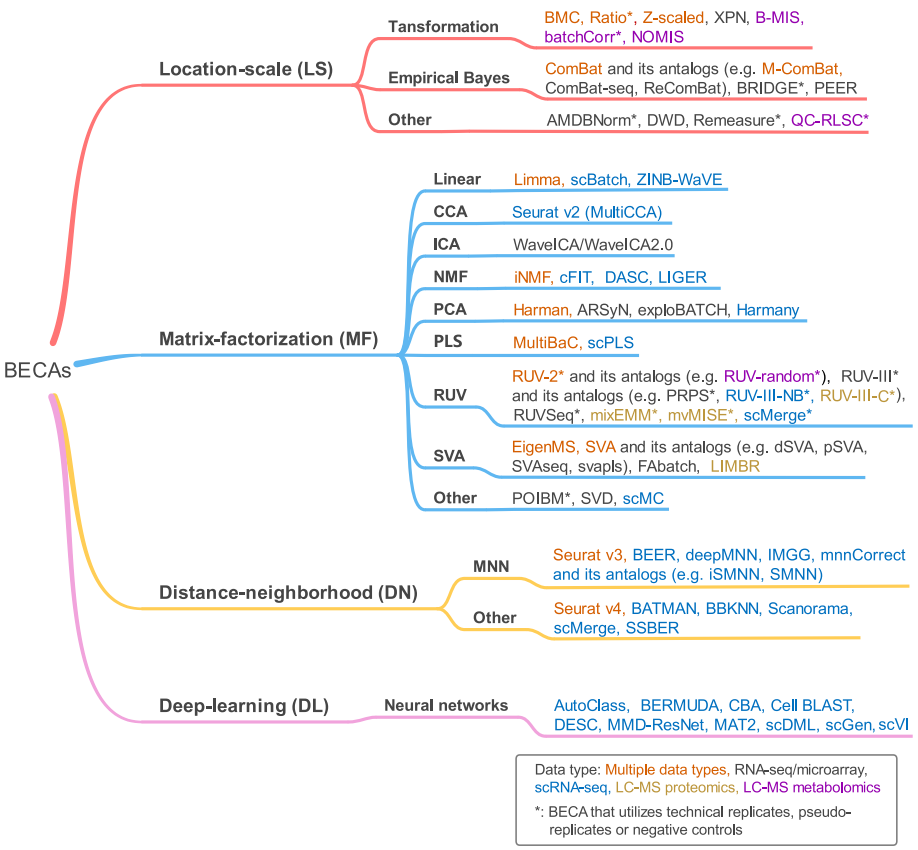

Fig. 3 Cluster tree of batch effect correction algorithms (BECAs). Detailed descriptions and references of BECAs are listed in additional file 1. The plot provides examples of representative BECAs, rather than an exhaustive compilation of all existing BECAs

design. It is not applicable when combining already-existing datasets, as the reference sample may not exist or be possible.

Matrix-factorization (MF) methods

MF methods assume that the observed data can be decomposed into a product of matrices that capture different sources of variation, and that this decomposition can be used to identify and remove batch effects from the data. Many matrix-factorization methods based on a diversity of statistical models have been developed and have been widely used in batch removal (Fig. 3).

The Surrogate Variable Analysis (SVA) method assumes that there are hidden factors, or surrogate variables, that are related to the batch effects and can be used to correct them [125]. SVA operates by specifying the number of latent factors to remove unwanted sources of variation while retaining differences among the specified primary variables. Alternatively, the software estimates the number of latent factors through a function call, and then performs the operation of estimating surrogate variables. Based on this algorithm, SVA can be successfully applied even when batch information is unclear. However, an inappropriate number of latent factors may result in the removal of potentially important biological information encoded in the latent variables. In this case,

SVA may not be appropriate for studies with unknown subgroups of biological interests [126], such as molecular subtyping studies. Several methods have been developed for improving the original SVA, such as direct SVA (dSVA) [127], permuted-SVA (pSVA) [128], and svapls [129]. Moreover, since SVA was initially developed based on microarray data, tools for adoption in RNA-seq (e.g., SVAseq [130]), proteomics (e.g., EigenMS [131], LIMBR [132]), and metabolomics (e.g., EigenMS [131, 133]) have been developed.

Remove Unwanted Variation (RUV) is a linear model-based batch correction algorithm that removes unwanted technical variation from gene expression data by first estimating unwanted variation using technical replicates or negative control genes. The unwanted variation is then subtracted from the original data to obtain corrected expression values. Traditional RUV methods include: RUVseq [134] and RUV-III [135] for RNA-seq, RUV-III-NB [136] for scRNA-seq, RUV-III-C [78] for LC-MS proteomics, and RUV-random [137] for LC-MS metabolomics. These methods require actual technical replicates or negative controls in the data to estimate unwanted variation. When such controls are not available, RUV-III-PRPS is developed to extend RUV-III algorithm by constructing pseudo-samples that mimic technical replicates [49]. Pseudo-samples are created by averaging gene expression levels within biological subpopulations that are homogeneous with respect to unwanted factors. RUV-III-PRPS then uses these pseudo-samples just like technical replicates to fit and remove unwanted variation.

Distance-neighborhood (DN) methods

DN methods assume that batch effects cause systematic differences between groups of samples that are close in the high-dimensional space or projected space, and that these differences can be corrected by adjusting the data to make these groups more similar.

BECAs based on mutual nearest neighbors (MNN), such as mnnCorrect [138], deepMNN [139], work by the removal of discrepancies between biologically related batches according to the presence of MNNs between batches, which are considered to define the most similar cells of the same type across batches. DN methods are effective in correcting for batch effects in scRNA-seq data because they rely on the assumptions, including (i) there is at least one cell population that is present in both batches, (ii) the batch effect is almost orthogonal to the biological subspace, and (iii) the batch effect is consistent across cells [138]. By identifying groups of cells with similar expression profiles and adjusting the data within each group, these methods can effectively correct for batch effects and improve the accuracy of downstream analyses [140]. In contrast, bulk RNA sequencing and proteomics data are often generated from a larger number of cells and are less prone to similar technical variations across samples due to differences in experimental conditions.

Deep-learning-based (DL) methods

DL methods usually use neural network algorithms to identify and remove batch effects from the data, including AutoClass [141], DESC [142], scGen [143], scVI [144], and so on. The basic assumption of these methods is to train a neural network to learn the relationship between gene expression values and experimental batch information in a dataset. The trained network can then be used to predict the batch information for each sample in the dataset based on its gene expression values. The predicted batch

information can then be used to adjust the gene expression values to correct for batch effects.

DL-based BECAs are often used in scRNA-seq because scRNA-seq data are high-dimensional and highly heterogeneous, which means they have a large number of samples with distinct gene expression profiles across multiple cell types. The relationships between gene expression profiles and batch information may be complex and non-linear. DL methods may learn complex nonlinear relationships between expressions and samples. Moreover, a large number of samples in scRNA-seq may provide sufficient data for training appropriate models and further obtaining satisfactory results. Of note, investigators should be aware of the risk of overfitting when DL-based BECAs are applied [141].

It is important to note that each type of BECAs has its own strengths and limitations, and the choice of method depends on the nature of the data, the sources of batch effects, and the specific goals of the analysis. Moreover, batch effect correction is still an active area of research, for example, with single-cell data. New methods are emerging and being evaluated. Therefore, it is recommended to carefully evaluate the performance of different BECAs in each specific context before choosing one for analysis.

Current consortium efforts

Many consortium efforts have been conducted to set standards and benchmark technologies, which also improve in batch evaluation and correction. In particular, consortium work is important and valuable for identifying the causes and sources of batch effects, developing and evaluating methods for reliable BECAs, and establishing best practices and guidelines for data analysis.

MAQC/SEQC

The MicroArray Quality Control (MAQC) and Sequencing Quality Control (SEQC) consortiums have made great efforts to assess the quality and reliability of emerging omics technologies, and to develop best practices for data analysis and interpretation [145]. The MAQC/SEQC projects have been conducted in four phases, namely MAQC-I, MAQC-II, MAQC-III/SEQC, and MAQC-IV/SEQC2. The MAQC-I project was published in 2006 and assessed the precision and comparability of microarray and quantitative RT-PCR datasets [146]. The MAQC-II was published in 2010 and assessed the performance of various machine-learning and data-analysis methods in microarray-based predictive models [115]. With the rapid development of RNA-seq technology, the MAQC-III/SEQC was published in 2014 and examined the reproducibility of RNA-seq and compared the performance of different RNA-seq platforms and DNA microarrays. Recently, the MAQC-IV/SEQC2 project was published in 2021 and benchmarked sequencing platforms in several applications [147], including genome sequencing [87, 148], cancer genomics [66, 149, 150], scRNA-seq [14], circulating tumor DNA [151], DNA methylation [152], and targeted RNA sequencing.

The MAQC consortium evaluated the impact of batch effects on gene expression measurements by analyzing the same set of RNA reference materials that were distributed to multiple laboratories around the world. It played an important role in highlighting the issue of batch effects in genomic data and in developing methods for batch

correction that have become standard practice in the field. Specifically, ratio-based expression profiles, defined as a fold-change or a ratio of expression levels between two sample groups for the same gene, agreed well across multiple transcriptomic technologies, including RNA-seq, microarray, and qPCR [73, 146]. Moreover, the ratio-based method was found to outperform others in terms of cross-batch prediction in clinical outcomes [36]. Furthermore, Risso et al. developed a new BECA strategy, called RUVseq, that adjusted for nuisance technical effects by performing factor analysis on suitable sets of control genes (e.g., ERCC spike-ins) or samples (e.g., replicate libraries) [134]. Recently, Chen et al. conducted a multi-center study focusing on the evaluation of data generation and bioinformatics tools using reference cell lines and found that batch-effect correction was by far the most important factor in correctly classifying the cells [14]. Nevertheless, reproducibility across centers was high when appropriate bioinformatic methods were applied. Additionally, RNA reference materials and datasets generated by the MAQC consortium have served as resources for the research community to develop and evaluate BECAs.

Multiomics Quartet project

The Quartet project team established a set of publicly available multiomics reference materials of matched DNA [86], RNA [69], proteins [82], and metabolites [83] derived from immortalized cell lines and assessed reliability across batches, labs, platforms, and omics types [13]. The results showed that the variation in gene expression measurements between laboratories was largely due to technical factors, such as differences in experimental protocols and equipment, rather than biological differences between samples [32]. Similar findings were obtained in DNA methylation, miRNA-seq, LC-MS proteomics, and LC-MS metablomics [13]. Importantly, the Quartet project found the "absolute" feature quantitation as the root cause of irreproducibility in multiomics measurement and data integration, and urged a paradigm shift from "absolute" to "ratio"-based multiomics profiling with common reference materials, i.e., by scaling the "absolute" omics data of study samples relative to those of concurrently measured universal reference materials on a feature-by-feature basis. The ratio-based multiomics data are much more resistant to batch effects [13].

Challenges and future directions

One of the major challenges of batch effect issues is evaluation and quantifying the impact of batch effects on the data. This can be difficult because batch effects can arise from various sources, including differences in sample preparation, sequencing technology, experimental conditions, and sometimes the sources are unknown or difficult to measure. It is important to develop methods that can accurately identify and quantify batch effects to minimize their impact on the downstream data analysis.

Another challenge in batch effect research is the generalization of batch effect correction methods across different datasets and experimental conditions. Batch effect correction methods may not always work well on new datasets or under different experimental conditions. On the other hand, because correcting batch effects requires fitting a model that captures the batch effects while preserving the biological signal, overfitting can occur and result in a loss of statistical power and generalization to new data. Hence, it

is important to develop methods that can generalize well across different datasets and experimental conditions to ensure the reliability and reproducibility of the data.

The third challenge is software and algorithm selection. There are various software packages and algorithms available for batch effect correction, each with different assumptions and limitations. Selecting the appropriate method for a specific dataset can be challenging, especially for non-experts.

Metagenomics (microbiome) research introduces unique challenges in batch effect correction due to its compositional structure and sparse count data [153, 154]. Traditional techniques like ComBat and RUV, developed for gene expression data, may not fully address these unique characteristics. Novel methods, such as ConQuR [155] and PLSDA [156], have been specifically designed for metagenomics data. Despite advancements, there is room for improvement in batch effect correction for microbiome data. For example, ConQuR's performance may be influenced by low-frequency taxa, and PLSDA may not be suitable when batch and treatment effects interact non-linearly. Thus, ongoing research focuses on developing more robust and efficient methods for batch effect correction in microbiome analysis.

Finally, batch effects are not limited to quantitative omics data alone, but also affect qualitative data, such as mutations, alternative splicing events, RNA editing events, and so on. However, there is a lack of established methods to remove batch effects from qualitative data. This highlights the need for developing and validating new methods to correct batch effects in qualitative data to ensure the accuracy and reproducibility of the results.

Conclusion

Batch effects are a common challenge in omics data analysis, especially in large-scale studies where samples are processed in batches or over an extended period of time. Assessing and mitigating batch effects is crucial for ensuring the reliability and reproducibility of omics data and minimizing the impact of technical variation on biological interpretation. As the data continue to grow, we expect experimental design and BECAs to also grow in importance and take center stage in large-scale applications in research and clinic. Quantifying multiomics data in a "ratio" scale during data generation stage has the potential to get rid of the enigmatic batch effects when common reference materials are adopted as the baseline in multiomics profiling.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03401-9.

Additional file 1. Detailed descriptions of representative BECAs.

Additional file 2. Review history.

Acknowledgements

We thank the Quartet Project team members, who contributed their time and resources to the design and implementation of this project.

Review history

The review history is available as Additional file 2.

Peer review information

Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
L.S., Y.Z., and Y.Y. conceived the review. Y.Y. and Y.M. constructed figures and tables. Y.Y., Y.Z., and L.S. wrote and/or revised the manuscript. All authors reviewed and approved the manuscript.

Funding
This study was supported in part by National Key R&D Project of China (2023YFC3402501 and 2023YFF0613302), the National Natural Science Foundation of China (T2425013, 32370701, 32170657, 32470692), Shanghai Municipal Science and Technology Major Project (2023SHZDZX02), State Key Laboratory of Genetic Engineering (SKLGE-2117), and the 111 Project (B13016).

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 11 July 2023 Accepted: 23 September 2024
Published online: 03 October 2024

References

1. Goh WWB, Yong CH, Wong L. Are batch effects still relevant in the age of big data? Trends Biotechnol. 2022;40:1029–40.
2. Cuklina J, Lee CH, Williams EG, Sajic T, Collins BC, Rodriguez Martinez M, et al. Diagnostics and correction of batch effects in large-scale proteomic studies: a tutorial. Mol Syst Biol. 2021;17:e10240.
3. Goh WWB, Wang W, Wong L. Why batch effects matter in omics data, and how to avoid them. Trends Biotechnol. 2017;35:498–507.
4. Lazar C, Meganck S, Taminau J, Steenhoff D, Coletta A, Molter C, et al. Batch effect removal methods for microarray gene expression data integration: a survey. Brief Bioinform. 2013;14:469–90.
5. Maceda I, Lao O. Analysis of the batch effect due to sequencing center in population statistics quantifying rare events in the 1000 genomes project. Genes (Basel). 2021;13:44.
6. Wickland DP, Ren Y, Sinnwell JP, Reddy JS, Pottier C, Sarangi V, et al. Impact of variant-level batch effects on identification of genetic risk factors in large sequencing studies. PLoS ONE. 2021;16:e0249305.
7. Anderson-Trocme L, Farouni R, Bourgey M, Kamatani Y, Higasa K, Seo JS, et al. Legacy data confound genomics studies. Mol Biol Evol. 2020;37:2–10.
8. Rasnic R, Brandes N, Zuk O, Linial M. Substantial batch effects in TCGA exome sequences undermine pan-cancer analysis of germline variants. BMC Cancer. 2019;19:783.
9. Mars RAT, Yang Y, Ward T, Houtti M, Priya S, Lekatz HR, et al. Longitudinal multi-omics reveals subset-specific mechanisms underlying irritable Bowel syndrome. Cell. 2020;183:1137–40.
10. Banchereau R, Hong S, Cantarel B, Baldwin N, Baisch J, Edens M, et al. Personalized immunomonitoring uncovers molecular networks that stratify lupus patients. Cell. 2016;165:1548–50.
11. Han W, Li L. Evaluating and minimizing batch effects in metabolomics. Mass Spectrom Rev. 2020;41:421–42.
12. Ugidos M, Nueda MJ, Prats-Montalban JM, Ferrer A, Conesa A, Tarazona S. MultiBaC: an R package to remove batch effects in multi-omic experiments. Bioinformatics. 2022;38:2657–8.
13. Zheng Y, Liu Y, Yang J, Dong L, Zhang R, Tian S, et al. Multi-omics data integration using ratio-based quantitative profiling with Quartet reference materials. Nat Biotechnol. 2023. https://doi.org/10.1038/s41587-41023-01934-41581.
14. Chen W, Zhao Y, Chen X, Yang Z, Xu X, Bi Y, et al. A multicenter study benchmarking single-cell RNA sequencing technologies using reference samples. Nat Biotechnol. 2021;39:1103–14.
15. Tran HTN, Ang KS, Chevrier M, Zhang X, Lee NYS, Goh M, et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol. 2020;21:12.
16. Freedman LP, Inglese J. The increasing urgency for standards in basic biologic research. Cancer Res. 2014;74:4024–9.
17. Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184:3573–3587 e3529.
18. Eddy S, Mariani LH, Kretzler M. Integrated multi-omics approaches to improve classification of chronic kidney disease. Nat Rev Nephrol. 2020;16:657–68.
19. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017;18:83.
20. Rosellini M, Marchetti A, Mollica V, Rizzo A, Santoni M, Massari F. Prognostic and predictive biomarkers for immunotherapy in advanced renal cell carcinoma. Nat Rev Urol. 2023;20:133–57.
21. Hassan M, Awan FM, Naz A, deAndres-Galiana EJ, Alvarez O, Cernea A, et al. Innovations in genomics and big data analytics for personalized medicine and health care: a review. Int J Mol Sci. 2022;23:4645.
22. Jiang P, Sinha S, Aldape K, Hannenhalli S, Sahinalp C, Ruppin E. Big data in basic and translational cancer research. Nat Rev Cancer. 2022;22:625–39.
23. Montaner J, Ramiro L, Simats A, Tiedt S, Makris K, Jickling GC, et al. Multilevel omics for the discovery of biomarkers and therapeutic targets for stroke. Nat Rev Neurol. 2020;16:247–64.
24. Li Y, Ma Y, Wang K, Zhang M, Wang Y, Liu X, et al. Using composite phenotypes to reveal hidden physiological heterogeneity in high-altitude acclimatization in a Chinese Han longitudinal cohort. Phenomics. 2021;1:3–14.

25. Xia Q, Thompson JA, Koestler DC. pwrBRIDGE: a user-friendly web application for power and sample size estimation in batch-confounded microarray studies with dependent samples. Stat Appl Genet Mol Biol. 2022;21:20220003.
26. Chen G, Ning B, Shi T. Single-cell RNA-seq technologies and related computational data analysis. Front Genet. 2019;10:317.
27. Yip SH, Sham PC, Wang J. Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data. Brief Bioinform. 2019;20:1583–9.
28. Phua SX, Lim KP, Goh WW. Perspectives for better batch effect correction in mass-spectrometry-based proteomics. Comput Struct Biotechnol J. 2022;20:4369–75.
29. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010;11:733–9.
30. Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray data. Biostatistics. 2012;13:539–52.
31. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8:118–27.
32. Yu Y, Zhang N, Mai Y, Chen Q, Cao Z, Chen Q, et al. Correcting batch effects in large-scale multiomic studies using a reference-material-based ratio method. Genome Biol. 2023;24:201.
33. Zhou W, Koudijs KKM, Bohringer S. Influence of batch effect correction methods on drug induced differential gene expression profiles. BMC Bioinformatics. 2019;20:437.
34. Li S, Labaj PP, Zumbo P, Sykacek P, Shi W, Shi L, et al. Detecting and correcting systematic variation in large-scale RNA sequencing data. Nat Biotechnol. 2014;32:888–95.
35. Li S, Tighe SW, Nicolet CM, Grove D, Levy S, Farmerie W, et al. Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. Nat Biotechnol. 2014;32:915–25.
36. Luo J, Schumacher M, Scherer A, Sanoudou D, Megherbi D, Davison T, et al. A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. Pharmacogenomics J. 2010;10:278–91.
37. Cardoso F, van't Veer LJ, Bogaerts J, Slaets L, Viale G, Delaloge S, et al. 70-gene signature as an aid to treatment decisions in early-stage breast cancer. N Engl J Med. 2016;375:717–29.
38. Lin S, Lin Y, Nery JR, Urich MA, Breschi A, Davis CA, et al. Comparison of the transcriptional landscapes between human and mouse tissues. Proc Natl Acad Sci U S A. 2014;111:17224–9.
39. Gilad Y, MizrahiMan O. A reanalysis of mouse ENCODE comparative gene expression data. F1000Res. 2015;4:121.
40. Mullard A. Half of top cancer studies fail high-profile reproducibility effort. Nature. 2021;600:368–9.
41. Errington TM, Mathur M, Soderberg CK, Denis A, Perfito N, Iorns E, et al. Investigating the replicability of preclinical cancer biology. Elife. 2021;10:e71601.
42. Baker M. 1,500 scientists lift the lid on reproducibility. Nature. 2016;533:452–4.
43. Freedman LP, Cockburn IM, Simcoe TS. The economics of reproducibility in preclinical research. PLoS Biol. 2015;13:e1002165.
44. Zhang S, Li X, Zhao S, Drobizhev M, Ai HW. Retraction note: a fast, high-affinity fluorescent serotonin biosensor engineered from a tick lipocalin. Nat Methods. 2021;18:575.
45. Yano Y, Mitoma N, Matsushima K, Wang F, Matsui K, Takakura A, et al. Retraction note: living annulative pi-extension polymerization for graphene nanoribbon synthesis. Nature. 2020;588:180.
46. Zhang S, Li X, Zhao S, Drobizhev M, Ai HW. A fast, high-affinity fluorescent serotonin biosensor engineered from a tick lipocalin. Nat Methods. 2021;18:258–61.
47. Errington TM, Denis A, Perfito N, Iorns E, Nosek BA. Challenges for assessing replicability in preclinical cancer biology. Elife. 2021;10:10.
48. Foox J, Tighe SW, Nicolet CM, Zook JM, Byrska-Bishop M, Clarke WE, et al. Performance assessment of DNA sequencing platforms in the ABRF next-generation sequencing study. Nat Biotechnol. 2021;39:1129–40.
49. Molania R, Foroutan M, Gagnon-Bartsch JA, Gandolfo LC, Jain A, Sinha A, et al. Removing unwanted variation from large-scale RNA sequencing data with PRPS. Nat Biotechnol. 2023;41:82–95.
50. Freedman LP, Venugopalan G, Wisman R. Reproducibility 2020: progress and priorities. F1000Res. 2017;6:604.
51. Wang C, Gong B, Bushel PR, Thierry-Mieg J, Thierry-Mieg D, Xu J, et al. The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. Nat Biotechnol. 2014;32:926–32.
52. Lippi G, Chance JJ, Church S, Dazzi P, Fontana R, Giavarina D, et al. Preanalytical quality improvement: from dream to reality. Clin Chem Lab Med. 2011;49:1113–26.
53. Su Y, Chen D, Yuan D, Lausted C, Choi J, Dai CL, et al. Multi-omics resolves a sharp disease-state shift between mild and moderate COVID-19. Cell. 2020;183(1479–1495):e1420.
54. Geyer PE, Holdt LM, Teupser D, Mann M. Revisiting biomarker discovery by plasma proteomics. Mol Syst Biol. 2017;13:942.
55. Halvey P, Farutin V, Koppes L, Gunay NS, Pappas DA, Manning AM, et al. Variable blood processing procedures contribute to plasma proteomic variability. Clin Proteomics. 2021;18:5.
56. Abraham RA, Agrawal PK, Acharya R, Sarna A, Ramesh S, Johnston R, et al. Effect of temperature and time delay in centrifugation on stability of select biomarkers of nutrition and non-communicable diseases in blood samples. Biochem Med (Zagreb). 2019;29:020708.
57. Jonasdottir HS, Brouwers H, Toes REM, Ioan-Facsinay A, Giera M. Effects of anticoagulants and storage conditions on clinical oxylipid levels in human plasma. Biochim Biophys Acta Mol Cell Biol Lipids. 2018;1863:1511–22.
58. Oddoze C, Lombard E, Portugal H. Stability study of 81 analytes in human whole blood, in serum and in plasma. Clin Biochem. 2012;45:464–9.
59. Xue VW, Ng SSM, Leung WW, Ma BBY, Cho WCS, Au TCC, et al. The effect of centrifugal force in quantification of colorectal cancer-related mRNA in plasma using targeted sequencing. Front Genet. 2018;9:165.

60. Wong SC, Ma BB, Lai PB, Ng SS, Lee JF, Hui EP, et al. The effect of centrifugation on circulating mRNA quantitation opens up a new scenario in expression profiling from patients with metastatic colorectal cancer. Clin Biochem. 2007;40:1277–84.
61. Zimmermann M, Traxler D, Simader E, Bekos C, Dieplinger B, Lainscak M, et al. In vitro stability of heat shock protein 27 in serum and plasma under different pre-analytical conditions: implications for large-scale clinical studies. Ann Lab Med. 2016;36:353–7.
62. Lippi G, Lima-Oliveira G, Brocco G, Bassi A, Salvagno GL. Estimating the intra- and inter-individual imprecision of manual pipetting. Clin Chem Lab Med. 2017;55:962–6.
63. Bobryk S, Goossen L. Variation in pipetting may lead to the decreased detection of antibodies in manual gel testing. Clin Lab Sci. 2011;24:161–6.
64. Pandya K, Ray CA, Brunner L, Wang J, Lee JW, DeSilva B. Strategies to minimize variability and bias associated with manual pipetting in ligand binding assays to assure data quality of protein therapeutic quantification. J Pharm Biomed Anal. 2010;53:623–30.
65. Ambardar S, Gupta R, Trakroo D, Lal R, Vakhlu J. High throughput sequencing: an overview of sequencing chemistry. Indian J Microbiol. 2016;56:394–404.
66. Xiao W, Ren L, Chen Z, Fang LT, Zhao Y, Lack J, et al. Toward best practice in cancer mutation detection with whole-genome and whole-exome sequencing. Nat Biotechnol. 2021;39:1141–50.
67. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nat Rev Genet. 2019;20:631–56.
68. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17:13.
69. Yu Y, Hou W, Wang H, Dong L, Liu Y, Sun S, et al. Quartet RNA reference materials improve the quality of transcriptomic data through ratio-based profiling. Nat Biotechnol. 2023. https://doi.org/10.1038/s41587-41023-01867-41589.
70. Mereu E, Lafzi A, Moutinho C, Ziegenhain C, McCarthy DJ, Alvarez-Varela A, et al. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. Nat Biotechnol. 2020;38:747–55.
71. Dal Molin A, Di Camillo B. How to design a single-cell RNA-sequencing experiment: pitfalls, challenges and perspectives. Brief Bioinform. 2019;20:1384–94.
72. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. The technology and biology of single-cell RNA sequencing. Mol Cell. 2015;58:610–20.
73. Su Z, Łabaj PP, Li S, Thierry-Mieg J, Thierry-Mieg D, Shi W, et al. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the sequencing quality control consortium. Nat Biotechnol. 2014;32:903–14.
74. Sprang M, Andrade-Navarro MA, Fontaine JF. Batch effect detection and correction in RNA-seq data using machine-learning-based automated assessment of quality. BMC Bioinformatics. 2022;23:279.
75. Li X, Zhang P, Wang H, Yu Y. Genes expressed at low levels raise false discovery rates in RNA samples contaminated with genomic DNA. BMC Genomics. 2022;23:554.
76. Sanchez-Illana A, Pineiro-Ramos JD, Sanjuan-Herraez JD, Vento M, Quintas G, Kuligowski J. Evaluation of batch effect elimination using quality control replicates in LC-MS metabolite profiling. Anal Chim Acta. 2018;1019:38–48.
77. Kuligowski J, Perez-Guaita D, Lliso I, Escobar J, Leon Z, Gombau L, et al. Detection of batch effects in liquid chromatography-mass spectrometry metabolomic data using guided principal component analysis. Talanta. 2014;130:442–8.
78. Poulos RC, Hains PG, Shah R, Lucas N, Xavier D, Manda SS, et al. Strategies to enable large-scale proteomics for reproducible research. Nat Commun. 2020;11:3793.
79. Jiang F, Liu Q, Li Q, Zhang S, Qu X, Zhu J, et al. Signal drift in liquid chromatography tandem mass spectrometry and its internal standard calibration strategy for quantitative analysis. Anal Chem. 2020;92:7690–8.
80. Brenes A, Hukelmann J, Bensaddek D, Lamond AI. Multibatch TMT reveals false positives, batch effects and missing values. Mol Cell Proteomics. 2019;18:1967–80.
81. Bell AW, Deutsch EW, Au CE, Kearney RE, Beavis R, Sechi S, et al. A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. Nat Methods. 2009;6:423–30.
82. Tian S, Zhan D, Yu Y, Liu M, Wang Y, Song L, et al. Quartet protein reference materials and datasets for multi-platform assessment of label-free proteomics. Genome Biol. 2023;24:202.
83. Zhang N, Chen Q, Zhang P, Zhou K, Liu Y, Wang H, et al. Quartet metabolite reference materials for inter-laboratory proficiency test and data integration of metabolomics profiling. Genome Biol. 2024;25:34.
84. Siskos AP, Jain P, Romisch-Margl W, Bennett M, Achaintre D, Asad Y, et al. Interlaboratory reproducibility of a targeted metabolomics platform for analysis of human serum and plasma. Anal Chem. 2017;89:656–65.
85. Goh WWB, Wong L. Advanced bioinformatics methods for practical applications in proteomics. Brief Bioinform. 2019;20:347–55.
86. Ren L, Duan X, Dong L, Zhang R, Yang J, Gao Y, et al. Quartet DNA reference materials and datasets for comprehensively evaluating germline variant calling performance. Genome Biol. 2023;24:270.
87. Pan B, Ren L, Onuchic V, Guan M, Kusko R, Bruinsma S, et al. Assessing reproducibility of inherited variants detected with short-read whole genome sequencing. Genome Biol. 2022;23:2.
88. Sahraeian SME, Mohiyuddin M, Sebra R, Tilgner H, Afshar PT, Au KF, et al. Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. Nat Commun. 2017;8:59.
89. Frohlich K, Brombacher E, Fahrner M, Vogele D, Kook L, Pinter N, et al. Benchmarking of analysis strategies for data-independent acquisition proteomics using a large-scale dataset comprising inter-patient heterogeneity. Nat Commun. 2022;13:2622.
90. O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med. 2013;5:28.
91. Ioannidis JP, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, et al. Increasing value and reducing waste in research design, conduct, and analysis. Lancet. 2014;383:166–75.

92. Ren L, Shi L, Zheng Y. Reference materials for improving reliability of multiomics profiling. Phenomics. 2024. https://doi.org/10.1007/s43657-023-00153-7. in press.
93. Sheng Q, Vickers K, Zhao S, Wang J, Samuels DC, Koues O, et al. Multi-perspective quality control of illumina RNA sequencing data analysis. Brief Funct Genomics. 2017;16:194–204.
94. Manimaran S, Selby HM, Okrah K, Ruberman C, Leek JT, Quackenbush J, et al. BatchQC: interactive software for evaluating sample and batch effects in genomic data. Bioinformatics. 2016;32:3836–8.
95. Liu X, Li N, Liu S, Wang J, Zhang N, Zheng X, et al. Normalization methods for the analysis of unbalanced transcriptome data: a review. Front Bioeng Biotechnol. 2019;7:358.
96. Wu Y, Li L. Sample normalization methods in quantitative metabolomics. J Chromatogr A. 2016;1430:80–95.
97. Quartet Project Team. Quartet RNA reference materials improve the quality of transcriptomic data through ratio-based profiling. Dataset. Open archive for miscellaneous data (OMIX). 2023. https://ngdc.cncb.ac.cn/omix/release/OMIX002254.
98. Quartet Project Team. Visualization of diagnsitics of batch effects. 2023. GitHub. https://doi.org/10.5281/zenodo.8101796.
99. van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learning Res. 2008;9:2579–605.
100. Diaz-Papkovich A, Anderson-Trocme L, Gravel S. A review of UMAP in population genetics. J Hum Genet. 2021;66:85–91.
101. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2019;37:38–44.
102. Kobak D, Berens P. The art of using t-SNE for single-cell transcriptomics. Nat Commun. 2019;10:5416.
103. Gandolfo LC, Speed TP. RLE plots: visualizing unwanted variation in high dimensional data. PLoS ONE. 2018;13:e0191629.
104. Bushel P. Principal variance component analysis. 2021. https://www.niehs.nih.gov/research/resources/software/biostatistics/pvca/index.cfm.
105. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36:411–20.
106. Varma S. Blind estimation and correction of microarray batch effect. PLoS ONE. 2020;15:e0231446.
107. Reese SE, Archer KJ, Therneau TM, Atkinson EJ, Vachon CM, de Andrade M, et al. A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal component analysis. Bioinformatics. 2013;29:2877–83.
108. Buttner M, Miao Z, Wolf FA, Teichmann SA, Theis FJ. A test metric for assessing single-cell RNA-seq batch correction. Nat Methods. 2019;16:43–9.
109. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019;16:1289–96.
110. Azizi E, Carr AJ, Plitas G, Cornish AE, Konopacki C, Prabhakaran S, et al. Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. Cell. 2018;174(1293–1308):e1236.
111. Hubert L, Arabie P. Comparing partitions. J Classif. 1985;2:193–218.
112. Batool F, Hennig C. Clustering with the average Silhouette width. Comput Stat Data Anal. 2021;158:107190.
113. Albrecht S, Sprang M, Andrade-Navarro MA, Fontaine JF. seqQscorer: automated quality control of next-generation sequencing data using machine learning. Genome Biol. 2021;22:75.
114. Parker HS, Leek JT. The practical effect of batch on genomic prediction. Stat Appl Genet Mol Biol. 2012;11:Article 10.
115. Shi L, Campbell G, Jones WD, Campagne F, Wen Z, Walker SJ, et al. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nat Biotechnol. 2010;28:827–38.
116. Handelman GS, Kok HK, Chandra RV, Razavi AH, Huang S, Brooks M, et al. Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods. AJR Am J Roentgenol. 2019;212:38–43.
117. Zhang Y, Jenkins DF, Manimaran S, Johnson WE. Alternative empirical Bayes models for adjusting for batch effects in genomic studies. BMC Bioinformatics. 2018;19:262.
118. Ni Z, Sun P, Zheng J, Wu M, Yang C, Cheng M, et al. JNK signaling promotes bladder cancer immune escape by regulating METTL3-mediated m6A modification of PD-L1 mRNA. Cancer Res. 2022;82:1789–802.
119. He YY, Xie XM, Zhang HD, Ye J, Gencer S, van der Vorst EPC, et al. Identification of hypoxia induced metabolism associated genes in pulmonary hypertension. Front Pharmacol. 2021;12:753727.
120. Nakayama M, Marchi H, Dmitrieva AM, Chakraborty A, Merl-Pham J, Hennen E, et al. Quantitative proteomics of differentiated primary bronchial epithelial cells from chronic obstructive pulmonary disease and control identifies potential novel host factors post-influenza A virus infection. Front Microbiol. 2022;13:957830.
121. Acharjee A, Hazeldine J, Bazarova A, Deenadayalu L, Zhang J, Bentley C, et al. Integration of metabolomic and clinical data improves the prediction of intensive care unit length of stay following major traumatic injury. Metabolites. 2021;12:12.
122. Stein CK, Qu P, Epstein J, Buros A, Rosenthal A, Crowley J, et al. Removing batch effects from purified plasma cell gene expression microarrays with modified ComBat. BMC Bioinformatics. 2015;16:63.
123. Adamer MF, Bruningk SC, Tejada-Arranz A, Estermann F, Basler M, Borgwardt K. reComBat: batch-effect removal in large-scale multi-source gene-expression data integration. Bioinform Adv. 2022;2:vbac071.
124. Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data.NAR. Genom Bioinform. 2020;2:lqaa078.
125. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007;3:1724–35.
126. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28:882–3.
127. Lee S, Sun W, Wright FA, Zou F. An improved and explicit surrogate variable analysis procedure by coefficient adjustment. Biometrika. 2017;104:303–16.

128. Parker HS, Leek JT, Favorov AV, Considine M, Xia X, Chavan S, et al. Preserving biological heterogeneity with a permuted surrogate variable analysis for genomics batch correction. Bioinformatics. 2014;30:2757–63.
129. Chakraborty S, Datta S, Datta S. svapls: an R package to correct for hidden factors of variability in gene expression studies. BMC Bioinformatics. 2013;14:236.
130. Leek JT. svaseq: removing batch effects and other unwanted noise from sequencing data. Nucleic Acids Res. 2014;42:e161.
131. Karpievitch YV, Taverner T, Adkins JN, Callister SJ, Anderson GA, Smith RD, et al. Normalization of peak intensities in bottom-up MS-based proteomics using singular value decomposition. Bioinformatics. 2009;25:2573–80.
132. Crowell AM, Greene CS, Loros JJ, Dunlap JC. Learning and imputation for mass-spec bias reduction (LIMBR). Bioinformatics. 2019;35:1518–26.
133. Karpievitch YV, Nikolic SB, Wilson R, Sharman JE, Edwards LM. Metabolomics data normalization with EigenMS. PLoS ONE. 2014;9:e116221.
134. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. Nat Biotechnol. 2014;32:896–902.
135. Molania R, Gagnon-Bartsch JA, Dobrovic A, Speed TP. A new normalization for Nanostring nCounter gene expression data. Nucleic Acids Res. 2019;47:6073–83.
136. Salim A, Molania R, Wang J, De Livera A, Thijssen R, Speed TP. RUV-III-NB: normalization of single cell RNA-seq data. Nucleic Acids Res. 2022;50:e96.
137. De Livera AM, Sysi-Aho M, Jacob L, Gagnon-Bartsch JA, Castillo S, Simpson JA, et al. Statistical methods for handling unwanted variation in metabolomics data. Anal Chem. 2015;87:3606–15.
138. Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol. 2018;36:421–7.
139. Zou B, Zhang T, Zhou R, Jiang X, Yang H, Jin X, et al. deepMNN: deep learning-based single-cell RNA sequencing data batch correction using mutual nearest neighbors. Front Genet. 2021;12:708981.
140. Wu Y, Zhang K. Tools for the analysis of high-dimensional single-cell RNA sequencing data. Nat Rev Nephrol. 2020;16:408–21.
141. Li H, Brouwer CR, Luo W. A universal deep neural network for in-depth cleaning of single-cell RNA-Seq data. Nat Commun. 2022;13:1901.
142. Li X, Wang K, Lyu Y, Pan H, Zhang J, Stambolian D, et al. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. Nat Commun. 2020;11:2338.
143. Lotfollahi M, Wolf FA, Theis FJ. scGen predicts single-cell perturbation responses. Nat Methods. 2019;16:715–21.
144. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018;15:1053–8.
145. Editorial Board. Sequencing benchmarked. Nat Biotechnol. 2021;39:1027.
146. Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol. 2006;24:1151–61.
147. Mercer TR, Xu J, Mason CE, Tong W. The Sequencing Quality Control 2 study: establishing community standards for sequencing in precision medicine. Genome Biol. 2021;22:306.
148. Reis ALM, Deveson IW, Madala BS, Wong T, Barker C, Xu J, et al. Using synthetic chromosome controls to evaluate the sequencing of difficult regions within the human genome. Genome Biol. 2022;23:19.
149. Fang LT, Zhu B, Zhao Y, Chen W, Yang Z, Kerrigan L, et al. Establishing community reference samples, data and call sets for benchmarking cancer mutation detection using whole-genome sequencing. Nat Biotechnol. 2021;39:1151–60.
150. Gong B, Li D, Kusko R, Novoradovskaya N, Zhang Y, Wang S, et al. Cross-oncopanel study reveals high sensitivity and accuracy with overall analytical performance depending on genomic regions. Genome Biol. 2021;22:109.
151. Deveson IW, Gong B, Lai K, LoCoco JS, Richmond TA, Schageman J, et al. Evaluating the analytical validity of circulating tumor DNA sequencing assays for precision oncology. Nat Biotechnol. 2021;39:1115–28.
152. Foox J, Nordlund J, Lalancette C, Gong T, Lacey M, Lent S, et al. The SEQC2 epigenomics quality control (EpiQC) study. Genome Biol. 2021;22:332.
153. Wang YW, Lêcao KA. Managing batch effects in microbiome data. Brief Bioinform. 2020;21:1954–70.
154. Fachrul M, Méric G, Inouye M, Pamp SJ, Salim A. Assessing and removing the effect of unwanted technical variations in microbiome data. Sci Rep. 2022;12:22236.
155. Ling WD, Lu JY, Zhao N, Lulla A, Plantinga AM, Fu WJ, et al. Batch effects removal for microbiome data via conditional quantile regression. Nat Commun. 2022;13:5418.
156. Wang YW, Cao KAL. PLSDA-batch: a multivariate framework to correct for batch effects in microbiome data. Brief Bioinform. 2023;24(2):bbac622.

Publisher's Note