Bioinformatics Advances, 2022, 1–7
https://doi.org/10.1093/bioadv/vbac071
Advance Access Publication Date: 6 October 2022
Original Paper

OXFORD

Software

reComBat: batch-effect removal in large-scale multi-source gene-expression data integration

Michael F. Adamer [1,2,*,†], Sarah C. Brüningk [1,2,†], Alejandro Tejada-Arranz[3], Fabienne Estermann[3], Marek Basler[3] and Karsten Borgwardt [1,2]

[1]Department of Biosystems Science and Engineering, ETH Zurich, Basel 4058, Switzerland, [2]Swiss Institute for Bioinformatics (SIB), Lausanne 1015, Switzerland and [3]Biozentrum, University of Basel, Basel 4056, Switzerland

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Thomas Lengauer

Received on August 11, 2022; revised on September 1, 2022; editorial decision on September 8, 2022; accepted on September 26, 2022

Abstract

Motivation: With the steadily increasing abundance of omics data produced all over the world under vastly different experimental conditions residing in public databases, a crucial step in many data-driven bioinformatics applications is that of data integration. The challenge of batch-effect removal for entire databases lies in the large number of batches and biological variation, which can result in design matrix singularity. This problem can currently not be solved satisfactorily by any common batch-correction algorithm.

Results: We present *reComBat*, a regularized version of the empirical Bayes method to overcome this limitation and benchmark it against popular approaches for the harmonization of public gene-expression data (both microarray and bulkRNAsq) of the human opportunistic pathogen *Pseudomonas aeruginosa*. Batch-effects are successfully mitigated while biologically meaningful gene-expression variation is retained. *reComBat* fills the gap in batch-correction approaches applicable to large-scale, public omics databases and opens up new avenues for data-driven analysis of complex biological processes beyond the scope of a single study.

Availability and implementation: The code is available at https://github.com/BorgwardtLab/reComBat, all data and evaluation code can be found at https://github.com/BorgwardtLab/batchCorrectionPublicData.

Contact: michael.adamer@bsse.ethz.ch

Supplementary information: Supplementary data are available at *Bioinformatics Advances* online.

1 Introduction

Data-driven computational biology greatly depends on the availability of large, integrated datasets to provide the necessary variety and statistical power for state-of-the-art (SOTA) machine and deep learning, as recently demonstrated by Alpha-Fold (Jumper *et al.*, 2021). In particular, an in-depth understanding of general trends in expression and transcription profiles are key for important research questions, such as overcoming microbial antibiotic resistance (Andersson *et al.*, 2020; Gil-Gil *et al.*, 2021), or cancer therapy failure (Kourou *et al.*, 2021; Malod-Dognin *et al.*, 2019). By mining large databases across studies, it may be possible to identify novel biological mechanisms that cannot be found by studying individual, small-scale experiments alone. This poses a problem shift toward the need for integrating diverse data obtained from numerous independent experiments.

Public databases, such as the Gene Expression Omnibus (GEO) (Barrett *et al.*, 2013; Edgar *et al.*, 2002), include independent studies collected over a large time span, under different biological and technical conditions. Hence, strong batch-effects (i.e. unwanted and biologically irrelevant variation) preclude a comprehensive analysis of pooled data and first need to be mitigated while desired biological variation [referred to in this article as '(experimental) design'] needs be retained.

Although a range of batch-correction algorithms has previously been suggested (Chazarra-Gil *et al.*, 2021; Lazar *et al.*, 2013; Rong *et al.*, 2020; Tran *et al.*, 2020), only a small subset of these remains applicable for this large-scale setting. In particular, most previous algorithms cannot incorporate high-dimensional experimental design information. Our goal for this study is to provide the community with a simple, yet effective extension of the popular and computationally efficient empirical Bayes method (Johnson *et al.*, 2007) (ComBat) to account for a large amount of highly correlated biological covariates. ComBat is based on ordinary linear regression and, therefore, will fail if the system is underdetermined.

We benchmark our method on simulated data and provide a real-world application in microarray and bulk RNAsq data, evaluating the impact of culture conditions on the gene-expression profiles of *Pseudomonas aeruginosa* (PA). PA is a Gram-negative bacterium

with a large genome (Stover *et al.*, 2000) that thrives in a variety of environments and has been declared a critical priority pathogen for the development of new antimicrobial treatments (Tacconelli *et al.*, 2018). A large range of studies have previously investigated the impact of culture conditions on the gene-expression profiles of PA. A comprehensive review of the perturbations caused by the microenvironmental cues is missing as a consequence of the lack of harmonized data allowing for a direct comparison.

The article is organized as follows. After reviewing relevant literature in Section 2, we introduce our *reComBat* algorithm (contribution i) in Section 3 as an extension of the ComBat algorithm to handle highly correlated covariates. In the second part of Section 3, we address the issue of assessing the efficacy of the batch-correction by introducing a large variety of evaluation metrics (contribution ii). In Section 4, we benchmark *reComBat* against a selection of SOTA batch-correction methods on simulated and real-world data. Finally, we present a large, harmonized dataset of PA expression profiles in response to different microenvironmental cues (contribution iii). We conclude Section 4 by demonstrating, as a proof of concept, the biological validity of the harmonized dataset. Section 5 comprises of a discussion and outlook.

2 Related work

A variety of batch-correction methods has previously been suggested for bulk and single-cell sequencing data [see e.g. Lazar *et al.* (2013), Tran *et al.* (2020) and Yu *et al.* (2021)]. Here, we focus on batch-correction of bulk data which can generally be divided into the following categories:

Normalization to reference genes or samples: Algorithms, such as cross-platform normalization (Shabalin *et al.*, 2008) or reference scaling (Kim *et al.*, 2007), which employ references, are infeasible in the public data domain: 'reference' or 'house keeping' genes do not exist for some organisms, particularly microbes, eliminating these as common ground for batch-effect correction. Given a large public dataset, overlapping samples or common reference experiments are unlikely.

Discretization methods: Approaches that discretize expression data into categories (e.g. 'expressed' versus 'not expressed') can be hard to implement rigorously without a relevant control. Furthermore, the information loss due to discretization may affect the results of any advanced downstream analysis of the harmonized data (McCall *et al.*, 2010; Warnat *et al.*, 2005).

Location-scale adjustments: These methods adjust the mean and/or variance of the genes, e.g. by standardization (Li and Wong, 2001) or batch mean-centering (Sims *et al.*, 2008). This only works if the batch-effect is a simple mean/variance shift and does not account for additional confounders. One of the most popular location-scale method is the empirical Bayes algorithm, ComBat (Johnson *et al.*, 2007). Despite reasonable success for the correction of local, i.e. within one experiment, or moderate (i.e. comprising few, biologically correlated) batch-effects most location-scale adjustment methods either provide insufficient correction in the presence of strong batch-effects (e.g. standardization) or are unable to account for highly correlated design features (e.g. ComBat).

Matrix factorization: This approach builds on decomposition, such as principal component analysis or singular value decomposition (Alter *et al.*, 2000) to identify and remove factors characterizing the batch. While this can work in small-scale experiments, it is unclear how to apply these methods when there is strong confounding of batch and biological variation. A tangential approach to matrix factorization is to estimate unwanted variation via surrogate variables (SVA) (Lazar *et al.*, 2013). Since in our setting, we assume that we know all sources of variation, we do not consider SVA.

Deep learning based: Recently, non-linear models, often based on neural/variational autoencoders or generative adversarial networks, have gained popularity [e.g. normAE (Rong *et al.*, 2020), AD-AE (Dincer *et al.*, 2020), scGen (Lotfollahi *et al.*, 2019) and Marouf *et al.* (2020)]. This class of models aims to find a batch-effect-free latent space representation of the data e.g. via adversarial training. While an advantage of these methods is their flexibility to account for batches, but also desired biological variation, a major drawback may be that the batch-effect is only removed in a low-dimensional latent space. Downstream analysis is necessarily constrained (Dincer *et al.*, 2020; Rong *et al.*, 2020). scGen is a notable exception as it provides a direct normalization at gene-expression level. However, large datasets are required and, in the absence of ground truth, the risk of overcorrection should be considered in addition to increased computational complexity.

3 Approach

In this section, we introduce the mathematical tools and start by defining our modification to the popular ComBat algorithm, *reComBat*, before introducing a range of possible evaluation metrics to gauge the efficacy of data harmonization.

3.1 Classical: ComBat

ComBat (Johnson *et al.*, 2007) is a well-established batch-correction algorithm employing a three-step process.

1. The gene expressions are estimated via an ordinary linear regression and the data are standardized.
2. The adjustment parameters are found by empirical Bayes estimates of parametric or non-parametric priors.
3. The standardized data are adjusted to remove the batch-effect.

The ComBat algorithm has seen many refinements and applications [see e.g. Čuklina *et al.* (2021), Müller *et al.* (2016) and Zhang *et al.* (2020)]. However, most datasets have been handling <20 data sources and did not come with an extensive design matrix. When the design matrix becomes large (many covariates) and sparse, unexpected issues can arise in Step 1 of the algorithm. To illustrate the classic algorithm, we use the slightly modified ansatz of Wachinger *et al.* (2021),

$$\boldsymbol{Y}_{ijk} = \underbrace{(\boldsymbol{X}\boldsymbol{\beta}^x)_{jk}}_{\text{desired variation}} + \underbrace{(\boldsymbol{C}\boldsymbol{\beta}^c)_{jk}}_{\text{undesired variation}} + \underbrace{\boldsymbol{\alpha}_k}_{\text{regression intercept}} + \underbrace{\boldsymbol{\beta}^g_{ik}}_{\text{additive batch-effect}} + \underbrace{\boldsymbol{\delta}_{ik}\epsilon_{ijk}}_{\text{multiplicative batch-effect}},$$

where $\boldsymbol{Y}_{ijk}$ is the gene expression of the *kth* gene in the *jth* sample of the *ith* batch. The matrices $\boldsymbol{X}$ and $\boldsymbol{C}$ are design matrices of desired and undesired variation with their corresponding matrices of regression coefficients $\boldsymbol{\beta}^x$ and $\boldsymbol{\beta}^c$. $\boldsymbol{\alpha}$ is a matrix of intercepts, and $\boldsymbol{\beta}^g$ and $\boldsymbol{\delta}$ parameterize the *additive* and *multiplicative* batch-effects. The tensor $\boldsymbol{\epsilon}$ is a 3D tensor of standard Gaussian random variables. Note, that we implicitly encode batch- and sample-dependency by dropping the relevant indices, i.e. $\boldsymbol{\beta}^g$ depends on the batch and gene, but is constant for each sample within the batch.

In the first step of the algorithm, the parameters $\boldsymbol{\beta}^x$, $\boldsymbol{\beta}^c$ and $\boldsymbol{\alpha}$ are fitted via an ordinary linear regression on

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}^x + \boldsymbol{C}\boldsymbol{\beta}^c + \boldsymbol{\alpha} = \tilde{\boldsymbol{X}}\boldsymbol{\beta}, \tag{2}$$

where $\tilde{\boldsymbol{X}} \in \mathbb{R}^{n\times m}$, where m is the number of features and n is the number of samples. Note that this formulation is equivalent to redefining $\boldsymbol{Y} \in \mathbb{R}^{n\times g}$, where g is the number of genes, and subsuming the batch and $\boldsymbol{C}$ features into $\tilde{\boldsymbol{X}}$. The intercept $\boldsymbol{\alpha}$ is inferred via the relation $\frac{1}{N}\sum_i n_i \boldsymbol{\beta}^g_{ik} = 0$ (Johnson *et al.*, 2007), where n_i is the number of samples in batch i, $\boldsymbol{\beta}_{ik}$ is the regression coefficient of batch i for gene k and N is the total number of samples. For ease of notation, in the remainder of this article, we will use this equivalent formulation.

Once, the model is fitted, the data are standardized, then the batch-effect parameters, $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\delta}}$ are estimated using a parametric or non-parametric empirical Bayes method. Finally, the data are adjusted. For details, please refer to the original publication (Johnson *et al.*, 2007).

3.2 Novel contribution: *reComBat*

Problem statement: Using standard results for ordinary linear regression, we know that if the matrix $A = \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$ is positive-definite, the optimization of (2) is strictly convex. However, if $\boldsymbol{A}$ is singular a

unique-solution the regression does not exist. Hence, if $\boldsymbol{A}$ is rank-deficient, i.e. the system is underdetermined, ComBat will not necessarily arrive at a unique-solution. Our goal in this work is to provide a computationally efficient solution for this problem to make the empirical Bayes method applicable also to large-scale public data harmonization.

Given the popularity of ComBat this issue does not seem to be encountered frequently. One possible explanation is that the sources of biological variation that are usually considered within the same experiment are limited and well-chosen. When integrating entire databases, however, the sources of biological variation are manifold and these can often only be encoded as categorical variables. One prominent example is considering all uploaded experimental data of a particular pathogen, which can result in hundreds of unique experimental conditions, some potentially highly correlated with other metadata. Encoding these as one-hot categorical variables creates a sparse, high-dimensional feature vector and, when many such categorical features are considered, then $m \approx n$. If, either $m > n$, or strong batch-design correlations exist, then, even for large-scale integration, $\boldsymbol{A}$ may be rank-deficient.

To mitigate this issue, we propose a modification of the estimation of gene-expression profiles by a linear model (Step 1 of the ComBat algorithm) by fitting the elastic net model—a standard approach from linear regression theory

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{x} + \boldsymbol{C}\hat{\boldsymbol{\beta}}^{c} + \hat{\boldsymbol{\alpha}}, \tag{3}$$

$$\hat{\boldsymbol{\beta}}^{x}, \hat{\boldsymbol{\beta}}^{c}, \hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\beta}^{x}, \boldsymbol{\beta}^{c}, \boldsymbol{\alpha}}{\operatorname{argmin}} [||\boldsymbol{Y} - \hat{\boldsymbol{Y}}||_2^2 + \lambda_1(||\boldsymbol{\beta}^{x}||_1 + ||\boldsymbol{\beta}^{c}||_1), \tag{4}$$

$$+\lambda_2(||\boldsymbol{\beta}^{x}||_2^2 + ||\boldsymbol{\beta}^{c}||_2^2)], \tag{5}$$

where $||\cdot||_p$ denotes the ℓ_p norm, and λ_1 and λ_2 are the LASSO and ridge regularization penalties. Due to this regularizing modification of the algorithm, we call our approach **re**gularized-ComBat, in short *reComBat*. Both, parameter fitting using the Empirical Bayes methods, and parameter adjustment on the standardized data follow the above outline for the ComBat algorithm. Note that *reComBat* essentially replaces a linear regression with a regularized regression and, hence, the increase of computational complexity of *reComBat* over ComBat is negligible.

The *reComBat* algorithm can be summarized in the following pseudo-code.

Algorithm 1 *reComBat*
Require: The data and the design: $Y, \tilde{X}$

1. Fit a regularized linear model: $Y = \tilde{X}\boldsymbol{\beta}$
2. Standardize Y
3. Obtain empirical Bayes estimates
4. Rescale Y: $Y \rightarrow \tilde{Y}$

Output: The corrected data: $\tilde{Y}$

4 Experiments

In this section, we apply *reComBat* to simulated and real-world microarray and bulkRNAsq data. We show quantitatively and qualitatively that *reComBat* is successful in removing substantial batch-effects while retaining biologically meaningful signal.

4.1 Evaluation metrics

A detailed description and definition of all evaluation metrics employed to score batch-correction efficacy is provided in Supplementary Material A. We included classifier-based [logistic regression-based balanced accuracy and F1-score, linear discriminant analysis (LDA) score], cluster-based (minimum separation number, cluster purity and Gini impurity) and sample distance-based [distance ratio score (DRS), Shannon entropy] metrics.

4.2 Experimental data

A detailed description is given in Supplementary Material B. Inspired by the graph theoretical notion of n-hop neighborhoods (Liu and Li, 2019), we group samples into so-called *Zero-Hops*. Each Zero-Hop defines a set of samples, which share the exact same experimental design. We first evaluate the approaches on synthetic data with singular design matrix and test a range of hyperparameter combinations for data generation [number of samples (100–2000), batches (3–100), design matrix features (3–20), relative disturbance size of metadata to batch (0.01–20), number of Zero-Hops, i.e. a set of samples sharing the experimental design (5–40)] and score run time, LDA score, Shannon entropy and cluster purity as a function thereof w.r.t. the ground truth. Additionally, data for 887 (114 batches, 39 Zero-Hops, see Supplementary Table S1) microarray and 340 bulkRNAsq samples (32 batches, 12 Zero-Hops, see Supplementary Table S2) were collected from the GEO, SRA and ENA data bases (Barrett *et al.*, 2013) with relevant metadata characterizing experimental design (culture conditions, PA strain). The obtained microarray design matrix is singular, whereas the RNA design matrix is not-singular, however, ill-conditioned. All input data were log-transformed.

4.3 Batch-correction methods

We tested our approach against a representative sample of baseline methods, in particular, standardization, marker gene elimination, principal component elimination, ComBat, Harmony (Korsunsky *et al.*, 2019) and scGen. Details on these methods can be found in the Supplementary Material C.

For *reComBat*, we used parametric priors for the empirical Bayes optimization and tested a variety of parameters including pure LASSO ($\lambda_2 = 0$), pure ridge ($\lambda_1 = 0$) and the full elastic net regression. The range of regularization strengths tested were all possible combinations [except for (0, 0)] of $\lambda_1 \in \{0, 10^{-2}, 10^{-1}, 1\}$ and $\lambda_2 \in \{0, 10^{-10}, \ldots, 10^{-1}, 1\}$. Note that smaller values of λ_1 yielded numerical instabilities.

4.4 Hyperparameter optimization results

A hyperparameter screen to optimize regularization strength and type on the default simulated, microarray and bulkRNAsq data yielded best results when ridge regression was used ($\lambda_1 = 0$) with $\lambda_2 \leq 0.001$ (see Supplementary Material D). The specific regularization parameter only had a minor influence and we continued with $\lambda_2 = 10^{-9}$ as an arbitrary choice. We observe that stronger, particularly LASSO, regularization achieves superior batch heterogeneity at the cost of decrease in Zero-Hop uniformity in real-world data. Notice that LASSO-*reComBat* performs implicit feature selection due the ℓ_1 regularization. This could hint to the fact that more balanced feature weighting (as provided by ridge-*reComBat*) is beneficial. In the following, we present results only for ridge-*reComBat*.

4.5 Evaluation on synthetic data

We benchmark *reComBat* on simulated data against popular batch-correction methods. Figure 1A and B shows the simulated ground truth distribution together with the distribution after applying batch-effects, and following data harmonization with *reComBat*. The ground truth results in terms of Zero-Hop clusters were qualitatively well reproduced by *reComBat*. Quantitative results in terms of LDA score difference to ground truth (see Supplementary Material E for Shannon entropy, Gini impurity and cluster purity) are shown in Figure 2A as a function of different data generation hyperparameters for the investigated correction methods. We observe that *reComBat* and scGen outperform Harmony and simple correction (PC or marker gene elimination, standardization). Notably, if scGen is trained with Zero-Hop labels its performance is greatly improved, however, also prone to overfitting. Overfitting

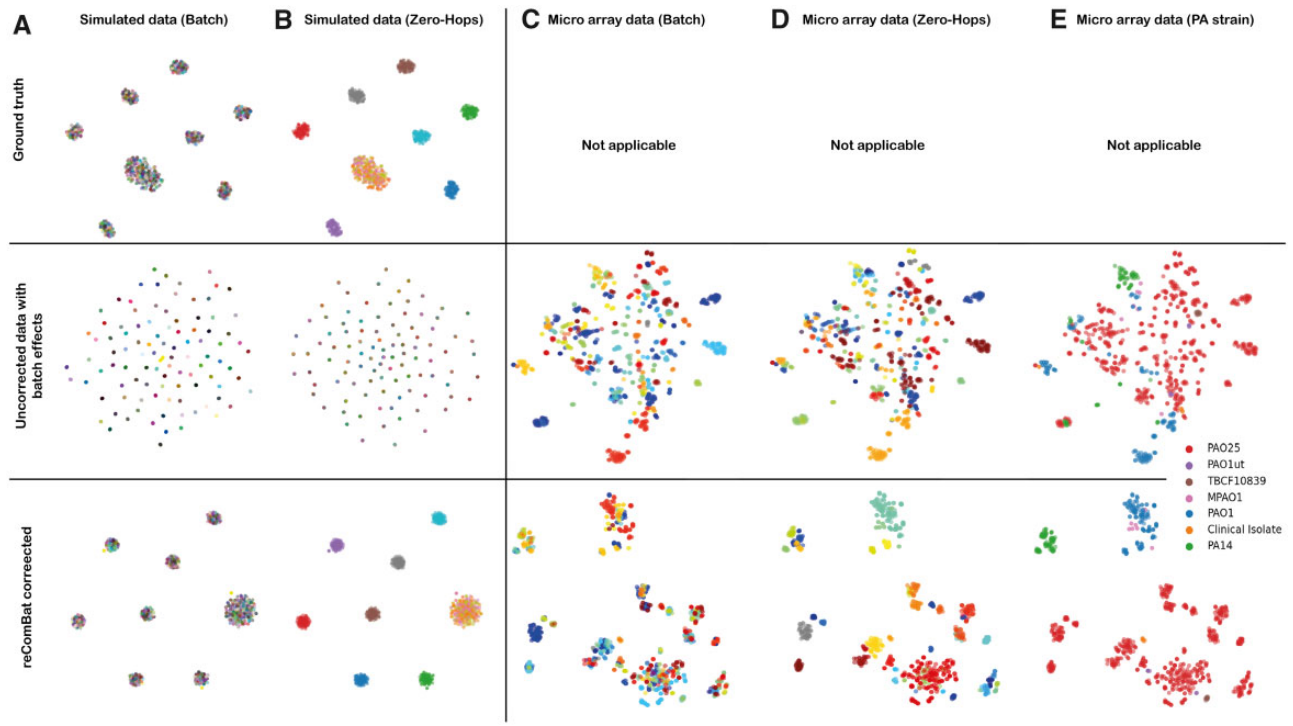

Fig. 1. *t*-SNE plots of the simulated (**A** and **B**) and microarray (**C** and **D**) datasets. For simulated data, we show ground truth (top), uncorrected (middle) and *reComBat* ($\lambda_1 = 0$, $\lambda_2 = 10^{-9}$) corrected (bottom) results. (Un)Corrected microarray data are colored by batches (top), Zero-Hops (middle) and microbial strain (bottom). Color scales do not reflect proximity of the relevant batches or Zero-Hops

was observed as positive LDA score differences for this method, indicating that a better LDA accuracy was obtained by scGen than possible based on the ground truth data. We only observe degradation of *reComBat* performance for smaller datasets of 100 samples (given 10 Zero-Hops). Run time was generally very quick and favorable for *reComBat* compared to Harmony, or scGen (trained on GPU).

4.6 Experimental benchmarking of *reComBat*

We show quantitatively and qualitatively that *reComBat* is successful in removing substantial batch-effects while retaining biologically meaningful signal in real-world data, too. Figure 1C and D gives an overview of the uncorrected and *reComBat* corrected microarray data colored by batch, Zero-Hops and microbial strain. Uncorrected data clusters by batch, indicating the presence of batch-effects, whereas clustering by biologically meaningful variation (e.g. by strain or Zero-Hop) is observed after correction. Additional overviews of *t*-SNE embeddings of batch-corrected expression data for all baseline models and data, colored by all design matrix elements are provided in Supplementary Material F.

We compared our baselines to the best performing *reComBat* model based on all evaluation metrics (Supplementary Material C) in Figure 2B. In terms of gauging the metrics themselves for the ability to detect batch-effects, we conclude that classifier-based metrics provide the clearest overview. Shannon entropy can detect a larger spread in batch versus Zero-Hop entropy, however, the findings may strongly vary by the specific subset. It can also be argued that entropy strongly depends on the choice of the number of nearest neighbors. Likewise, the median pairwise distance and DRS metrics show some ability to detect batch-correction, but due to the strong dependency on the individual Zero-Hop the spread in values may be large. The minimum separation clustering clearly shows when a batch-correction can be considered effective. However, due to repeated clustering, calculation of minimum separation number is computationally far more expensive than distance-based metrics. A good mid-point between classifier- and cluster-based evaluations are cluster-purity measures, which show good resolution and manageable dependency on the Zero-Hop.

Data standardization, and marker gene elimination only had a minor, insignificant (all Mann–Whitney *U*-Test *P*-values >0.05) effect when compared to the raw data, independent of the underlying metric and dataset. Despite, markedly different results compared to the uncorrected baseline, Harmony could not achieve sufficient batch-correction characterized by poor performance in classifier and cluster-based metrics throughout. We suggest that the large number of design matrix elements and comparably strong batch-effect could lead to this result. Importantly, *reComBat* achieved good scores throughout all evaluation metrics for all datasets (bulkRNAsq given in Supplementary Material), whereas performance of other correction methods, such as PC elimination, scGen and ComBat, varied depending on data and metric. As expected, singularity of the design matrix led to poor performance of ComBat (microarray data), whereas bulkRNAsq data with a non-singular design matrix achieved the best results for this method. For scGen it was key to provide information on Zero-Hops as labels to the algorithm [scGen (Zero-Hop)], whereas simply relying on design matrix covariates led to poor correction. Such behavior may complicate applications where specific label information may not be available in practice. Label construction based on a large design matrix may not always be straightforward and label-free correction methods, such as *reComBat* would be at an advantage.

4.7 Characterization of the harmonized microarray dataset

In order to preclude overcorrection (Zindler *et al.*, 2020) in the absence of ground truth, we demonstrate that biologically meaningful expression profiles are retained after batch-correction. As representative examples, we analyzed data subsets by oxygenation status, culture medium richness, growth phase, or clinical versus laboratory PA strains in our microarray dataset (Supplementary Material G). We identified Zero-Hop marker genes driving the differences between selected pairwise comparisons and assessed their relevance to underlying biological pathways. Pathways previously known to be

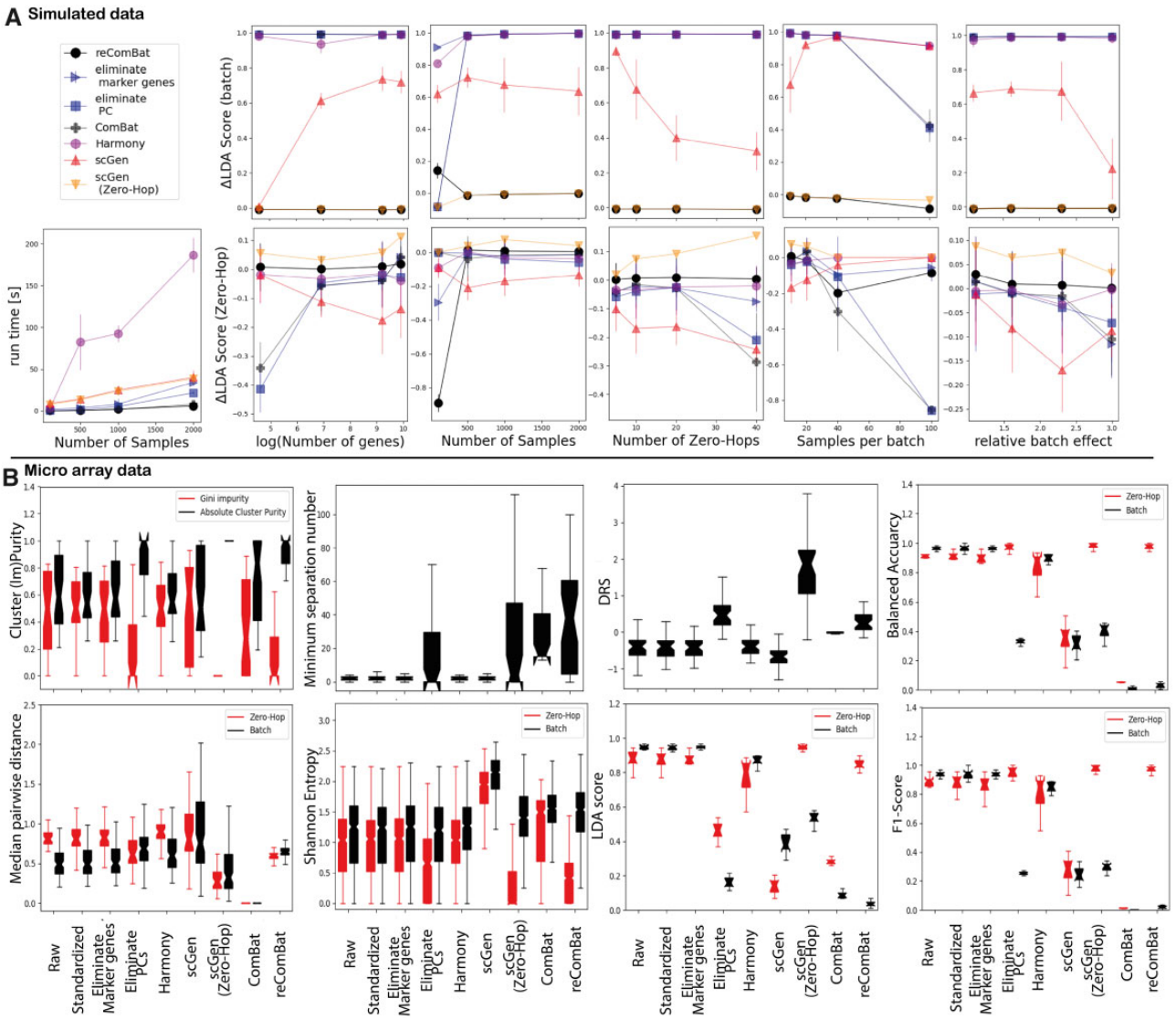

Fig. 2. (**A**) Overview over results based on different simulated datasets scored in terms of LDA score difference to ground truth for batch and Zero-Hops. Results represent mean values and standard deviations over 10 independent repeats. (**B**) Evaluation metrics scoring the impact of batch-effects by evaluating the variety of different batches and/or Zero-Hops of the (un-) corrected microarray dataset. Box plots represent the lower and upper quartiles (box) together with the median (central dents) and full range (whiskers) over all samples, clusters or Zero-Hops depending on the relevant metric. LDA scores and LR classification performance are reported over 10 cross-validation folds

important in the relevant culture conditions were identified. For instance, when comparing standard to hypoxic conditions, we found that genes involved in aerotaxis (Hong *et al.*, 2004), Fe-S cluster biogenesis (Romsang *et al.*, 2015) and iron acquisition (Glanville *et al.*, 2021; Hannauer *et al.*, 2012) are major drivers of differences. When comparing cultures in exponential to stationary phase under hypoxia conditions, genes involved in pyoverdin (Drake *et al.*, 2007; Vandenende *et al.*, 2004) and pyochelin (Ankenbauer and Quan, 1994; Reimmann *et al.*, 2001) biosynthesis and transport, iron starvation (Alontaga *et al.*, 2009; Hassett *et al.*, 1997; Zhao and Poole, 2000) and quorum sensing (Kim *et al.*, 2012) were relevant. Finally, for a comparison between the laboratory strain PAO1 versus clinical isolates, we found cup genes (PA4081-PA4084, PA0994) that are involved in motility and attachment in biofilm formation (Ruer *et al.*, 2007). This indicates a difference in attachment between those strains that might be coming from the environment the strains have adapted to grow in (laboratory versus patient). In all cases, a large amount of hypothetical genes of unknown function also flagged up – an expected observation as roughly two-thirds of the genes encoded in the PA genome have an unknown function. The harmonized dataset hence serves for hypothesis generation motivating further (experimental) validation.

5 Discussion

Public databases play an increasingly important role for data-driven meta-analysis in computational biology. Despite great efforts to harmonize data collection, considerable, yet unavoidable, biological/technical variation may mask true signal if data are pooled from several sources. To draw generalizable conclusions from agglomerated data, it is essential to correct such batch-effects in a setting where overlapping samples, or standardized controls, are unavailable. When large numbers of (>20) batches coincide with desired biological variation, a range of standard batch-correction algorithms are inapplicable. We would like to stress that this evaluation scenario greatly differs from previously analyzed batch-correction settings where comparably few (2–5) batches with large number of overlapping samples were included, or comparably small batch-effects within a single study were corrected (Tran *et al.*, 2020). A key assumption of meta-analysis of published data is the coincidence

of 'batch' with 'study'. Given the substantial manual data curation to extract relevant design matrix information for experimental data the variety of data types (microarray and bulkRNAsq) and organisms (PA) assessed in addition to simulated data was limited. *reComBat* is a simple yet effective, means of mitigating highly correlated experimental conditions through regularization and we compared various elastic net regularization strengths for the purpose of meta-analysis based on large-scale public data. We note that given the large number of batch-correction methods available, we only included representative examples for key concepts, including deep, non-linear models (scGen), Harmony, marker gene and PC elimination to benchmark our linear empirical Bayes method.

In case of a singular design matrix *reComBat* outperformed standard approaches, including data standardization, PC and marker gene elimination, Harmony and scGen if no additional information regarding the evaluation endpoints (here Zero-Hops) was given to either of the methods. We demonstrate not only the superiority of *reComBat* compared to these baselines but, by providing a large variety of evaluation metrics, also give a notion of overall performance.

Importantly, in any large-scale meta-analysis setting, a ground truth is unavailable. Here, biological validation is essential prior to hypothesis generation and we demonstrate this for *reComBat*. Due to this fact, we excluded some popular deep models [e.g. normAE (Rong *et al.*, 2020) and AD-AE (Dincer *et al.*, 2020)] from this study as they only provide a latent representation rather than direct correction at gene-expression level. These methods would likely provide good batch-correction, however, downstream analysis via e.g. differential gene expression is impossible. There is also growing concern that batch-correction, particularly deep models, may overcorrect and remove biological signal. Although synthetic data addresses this challenge, algorithm performance varies between use-cases and the risk of overcorrection persists. We demonstrate this based on scGen (Zero-Hop) in our benchmark. Both scGen and Harmony (in the published python packages) do not allow for a separation of batch-correction training and validation to test for overfitting by cross-validation—*reComBat* indeed could be used in a cross-validation setting. Notably, in case of e.g. large-scale single-cell RNA sequencing, the situation may in fact be favorable for non-linear approaches—which is not the setting of interest here.

It was possible to show that *reComBat* retained biologically meaningful target pathways identified in a literature-based validation. By mining the harmonized dataset, we can now perform comparisons that have, to the best of our knowledge, never been directly performed before for the purpose of hypothesis generation. For instance, when we compare growth in LB with growth in media that have fewer nutrients, we find that several nutrient (Bains *et al.*, 2012; Ball *et al.*, 2002; Faure *et al.*, 2014; Jones *et al.*, 2021; Lewenza *et al.*, 2011; Quesada *et al.*, 2016) and metal (Alontaga *et al.*, 2009; Merriman *et al.*, 1995) uptake pathways are deferentially regulated. Experimental validation of the proposed findings is a key in confirming information on the underlying biological mechanisms.

With >5000 citations, ComBat is one of the most popular batch-correction methods today applied to a large variety of data types and organisms (Wachinger *et al.*, 2021). In this study, we showed how an adaptation of this popular algorithm can drastically increase its usability. ComBat benefits from low computational cost, rigorous underlying theory, interpretability and is easy to apply in practice. We specifically want to recommend *reComBat* in a setting of comparably strong batch-effects and diverse experimental designs as are frequently observed within publicly sourced data from different laboratories. We acknowledge the small methodological differences between ComBat and *reComBat* but stress the importance of this adaptation to make a well-established method suitable for large-scale public data integration. By publishing *reComBat* as a python package (https://github.com/BorgwardtLab/reComBat) our method is readily available to the community. We also make the harmonized datasets with their metadata available to the wider research community (https://github.com/BorgwardtLab/batchCorrectionPublicData).

6 Conclusion

We have addressed the challenge of harmonizing large, and highly diverse public data for downstream meta-analysis. Aiming at high community acceptance and a computationally efficient solution, we extend the well-established ComBat algorithm through the addition of regularization. We evaluate our novel algorithm on simulated, and public microarray and bulkRNAsq data. A variety of evaluation metrics attest comparable, or superior correction of batch-effects as established baseline models. Our analysis constitutes a proof of principle to motivate and enable further large-scale meta-analyses.

Acknowledgements

We thank Dr Robert Ivanek (DBM Bioinformatics Core Facility at the University Basel and Swiss Institute of Bioinformatics) for advice on downloading and processing of the datasets.

Author contributions

M.F.A., S.C.B., M.B. and K.B. designed the project. M.F.A. implemented the reComBat python package and synthetic data simulation. S.C.B. performed experiments, benchmarks, and created visualizations. F.E. extracted microarray data from the Gene Expression Omnibus, including manual extraction of metadata information. A.T.-A. extracted and preprocessed RNAsq data, including manual extraction of metadata information. F.E. and A.T.-A. characterized the harmonized dataset. M.B. and K.B. coordinated the project. All authors contributed to the writing of the manuscript.

Funding

This work was supported by the National Center of Competence in Research AntiResist funded by the Swiss National Science Foundation [51NF40_180541]; and funded in part by the Alfried Krupp Prize for Young University Teachers of the Alfried Krupp von Bohlen und Halbach-Stiftung to K.B.

Conflict of Interest: none declared.

Data availability

reComBat is provided as a python (https://github.com/BorgwardtLab/reComBat). The harmonized datasets with their metadata are available on https://github.com/BorgwardtLab/batchCorrectionPublicData. All individual data sets have previously been published on the GEO, SRA and ENA data bases.

References

Alontaga,A.Y. *et al.* (2009) Structural characterization of the hemophore HasAp from *Pseudomonas aeruginosa*: NMR spectroscopy reveals protein-protein interactions between Holo-HasAp and hemoglobin. *Biochemistry*, **48**, 96–109.

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

Andersson,D.I. *et al.* (2020) Antibiotic resistance: turning evolutionary principles into clinical reality. *FEMS Microbiol. Rev.*, **44**, 171–188.

Ankenbauer,R.G. and Quan,H.N. (1994) FptA, the Fe(III)-pyochelin receptor of *Pseudomonas aeruginosa*: a phenolate siderophore receptor homologous to hydroxamate siderophore receptors. *J. Bacteriol.*, **176**, 307–319.

Bains,M. *et al.* (2012) Phosphate starvation promotes swarming motility and cytotoxicity of *Pseudomonas aeruginosa*. *Appl. Environ. Microbiol.*, **78**, 6762–6768.

Ball,G. *et al.* (2002) A novel type II secretion system in *Pseudomonas aeruginosa*. *Mol. Microbiol.*, **43**, 475–485.

Barrett,T. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.

Chazarra-Gil,R. *et al.* (2021) Flexible comparison of batch correction methods for single-cell RNA-seq using BatchBench. *Nucleic Acids Res.*, **49**, e42.

Čuklina,J. *et al.* (2021) Diagnostics and correction of batch effects in large-scale proteomic studies: a tutorial. *Mol. Syst. Biol.*, **17**, e10240.
Dincer,A.B. *et al.* (2020) Adversarial deconfounding autoencoder for learning robust gene expression embeddings. *Bioinformatics*, **36**, i573–i582.
Drake,E.J. *et al.* (2007) The 1.8 Å crystal structure of PA2412, an MbtH-like protein from the pyoverdine cluster of *Pseudomonas aeruginosa*. *J. Biol. Chem.*, **282**, 20425–20434.
Edgar,R. *et al.* (2002) Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
Faure,L.M. *et al.* (2014) Characterization of a novel two-partner secretion system implicated in the virulence of *Pseudomonas aeruginosa*. *Microbiology (Reading)*, **160**, 1940–1952.
Gil-Gil,T. *et al.* (2021) Antibiotic resistance: time of synthesis in a post-genomic age. *Comput. Struct. Biotechnol. J.*, **19**, 3110–3124.
Glanville,D.G. *et al.* (2021) A high-throughput method for identifying novel genes that influence metabolic pathways reveals new iron and heme regulation in *Pseudomonas aeruginosa*. *mSystems*, **6**, 1.
Hannauer,M. *et al.* (2012) The PvdRT-OpmQ efflux pump controls the metal selectivity of the iron uptake pathway mediated by the siderophore pyoverdine in *Pseudomonas aeruginosa*. *Environ. Microbiol.*, **14**, 1696–1708.
Hassett,D.J. *et al.* (1997) Fumarase C activity is elevated in response to iron deprivation and in mucoid, alginate-producing *Pseudomonas aeruginosa*: cloning and characterization of fumC and purification of native fumC. *J. Bacteriol.*, **179**, 1442–1451.
Hong,C.S. *et al.* (2004) Chemotaxis proteins and transducers for aerotaxis in *Pseudomonas aeruginosa*. *FEMS Microbiol. Lett.*, **231**, 247–252.
Johnson,W.E. *et al.* (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**, 118–127.
Jones,R.A. *et al.* (2021) Phosphorus stress induces the synthesis of novel glycolipids in *Pseudomonas aeruginosa* that confer protection against a last-resort antibiotic. *ISME J.*, **15**, 3303–3314.
Jumper,J. *et al.* (2021) Highly accurate protein structure prediction with alphafold. *Nature*, **596**, 583–589.
Kim,K.-Y. *et al.* (2007) An attempt for combining microarray data sets by adjusting gene expressions. *Cancer Res. Treat.*, **39**, 74–81.
Kim,S.-K. *et al.* (2012) AntR-mediated bidirectional activation of antA and antR, anthranilate degradative genes in *Pseudomonas aeruginosa*. *Gene*, **505**, 146–152.
Korsunsky,I. *et al.* (2019) Fast, sensitive and accurate integration of single-cell data with harmony. *Nat. Methods*, **16**, 1289–1296.
Kourou,K. *et al.* (2021) Applied machine learning in cancer research: a systematic review for patient diagnosis, classification and prognosis. *Comput. Struct. Biotechnol. J.*, **19**, 5546–5555.
Lazar,C. *et al.* (2013) Batch effect removal methods for microarray gene expression data integration: a survey. *Brief. Bioinform.*, **14**, 469–490.
Lewenza,S. *et al.* (2011) The olsA gene mediates the synthesis of an ornithine lipid in *Pseudomonas aeruginosa* during growth under phosphate-limiting conditions, but is not involved in antimicrobial peptide susceptibility. *FEMS Microbiol. Lett.*, **320**, 95–102.
Li,C. and Wong,W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA*, **98**, 31–36.
Liu,W. and Li,Z. (2019) An efficient parallel algorithm of n-hop neighborhoods on graphs in distributed environment. *Front. Comput. Sci.*, **13**, 1309–1325.
Lotfollahi,M. *et al.* (2019) scGen predicts single-cell perturbation responses. *Nat. Methods*, **16**, 715–721.
Malod-Dognin,N. *et al.* (2019) Towards a data-integrated cell. *Nat. Commun.*, **10**, 805.
Marouf,M. *et al.* (2020) Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks. *Nat. Commun.*, **11**, 166.
McCall,M.N. *et al.* (2010) Frozen robust multiarray analysis (fRMA). *Biostatistics*, **11**, 242–253.
Merriman,T.R. *et al.* (1995) Nucleotide sequence of pvdD, a pyoverdine biosynthetic gene from *Pseudomonas aeruginosa*: pvdD has similarity to peptide synthetases. *J. Bacteriol.*, **177**, 252–258.
Müller,C. *et al.* (2016) Removing batch effects from longitudinal gene Expression - Quantile normalization plus ComBat as best approach for microarray transcriptome data. *PLoS One*, **11**, e0156594.
Quesada,J.M. *et al.* (2016) The activity of the *Pseudomonas aeruginosa* virulence regulator σVreI is modulated by the anti-σ factor VreR and the transcription factor PhoB. *Front. Microbiol.*, **7**, 1159.
Reimmann,C. *et al.* (2001) Essential PchG-dependent reduction in pyochelin biosynthesis of *Pseudomonas aeruginosa*. *J. Bacteriol.*, **183**, 813–820.
Romsang,A. *et al.* (2015) *Pseudomonas aeruginosa* IscR-Regulated ferredoxin NADP(+) reductase gene (fprB) functions in Iron-Sulfur cluster biogenesis and multiple stress response. *PLoS One*, **10**, e0134374.
Rong,Z. *et al.* (2020) NormAE: deep adversarial learning model to remove batch effects in liquid chromatography mass spectrometry-based metabolomics data. *Anal. Chem.*, **92**, 5082–5090.
Ruer,S. *et al.* (2007) Assembly of fimbrial structures in *Pseudomonas aeruginosa*: functionality and specificity of chaperone-usher machineries. *J. Bacteriol.*, **189**, 3547–3555.
Shabalin,A.A. *et al.* (2008) Merging two gene-expression studies via cross-platform normalization. *Bioinformatics*, **24**, 1154–1160.
Sims,A.H. *et al.* (2008) The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets – improving meta-analysis and prediction of prognosis. *BMC Med. Genomics*, **1**, 42.
Stover,C.K. *et al.* (2000) Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. *Nature*, **406**, 959–964.
Tacconelli,E. *et al.*; WHO Pathogens Priority List Working Group. (2018) Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *Lancet Infect. Dis.*, **18**, 318–327.
Tran,H.T.N. *et al.* (2020) A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.*, **21**, 12.
Vandenende,C.S. *et al.* (2004) Functional characterization of an aminotransferase required for pyoverdine siderophore biosynthesis in *Pseudomonas aeruginosa* PAO1. *J. Bacteriol.*, **186**, 5596–5602.
Wachinger,C. *et al.*; Alzheimer's Disease Neuroimaging Initiative. (2021) Detect and correct bias in multi-site neuroimaging datasets. *Med. Image Anal.*, **67**, 101879.
Warnat,P. *et al.* (2005) Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics*, **6**, 265–215.
Yu,X. *et al.* (2021) *Statistical and Bioinformatics Analysis of Data from Bulk and Single-Cell RNA Sequencing Experiments*. Springer, New York, NY, pp. 143–175.
Zhang,Y. *et al.* (2020) ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom. Bioinform.*, **2**, lqaa078.
Zhao,Q. and Poole,K. (2000) A second tonB gene in *Pseudomonas aeruginosa* is linked to the exbB and exbD genes. *FEMS Microbiol. Lett.*, **184**, 127–132.
Zindler,T. *et al.* (2020) Simulating ComBat: how batch correction can lead to the systematic introduction of false positive results in DNA methylation microarray studies. *BMC Bioinformatics*, **21**, 271.