reComBat: Batch effect removal in large-scale, multi-source omics data integration

Michael F. Adamer*[1,2][0000−0002−8996−7167], Sarah C. Brüningk*[1,2][0000−0003−3176−1032], Alejandro Tejada-Arranz[3][0000−0002−8452−8245], Fabienne Estermann[3][0000−0002−9629−4386], Marek Basler[3][0000−0001−5414−2088], and Karsten M. Borgwardt[1,2][0000−0001−7221−2393]

* These authors share first authorship

[1] Machine Learning & Computational Biology Lab, Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland

[2] Swiss Institute for Bioinformatics (SIB), Lausanne, Switzerland

[3] Biozentrum, University of Basel, Basel, Switzerland

michael.adamer@bsse.ethz.ch

Abstract. With the steadily increasing abundance of omics data produced all over the world, sometimes decades apart and 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 and coincide of both batches and desired, biological variation resulting in design matrix singularity. This problem currently cannot be solved by any common batch correction algorithm. In this study, we present *reComBat*, a regularised version of the empirical Bayes method to overcome this limitation. We demonstrate our approach for the harmonisation of public gene expression data of the human opportunistic pathogen *Pseudomonas aeruginosa* and study a several metrics to empirically demonstrate that 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.

Keywords: batch correction, combat, empirical Bayes, batch effects, pseudomonas aeruginosa, microarray

1 Introduction

Data-driven computational biology greatly depends on the availability of large, integrated data sets to provide the necessary variety and statistical power for state-of-the-art machine and deep learning approaches, as recently demonstrated by Alpha-Fold [24]. In particular, an in-depth understanding of general trends in expression and transcription profiles are key to further the progress on important research questions such as overcoming microbial antibiotic resistance, [16, 3] or cancer therapy failure [27, 33]. By mining large databases of individual experiments, it may be possible to identify novel biological mechanisms that cannot be found by studying each individual, small-scale experiment alone. This possibility poses a problem shift towards the need for integrating more biologically diverse data obtained from numerous independent experiments, rather than drawing direct experimental comparisons.

To this end, public databases such as the Gene Expression Omnibus (GEO) [7, 12], are essential data sources. However, as the published independent studies originate from different laboratories, are collected over a large time span, under different biological, and technical conditions, strong batch effects (i.e. unwanted and biologically irrelevant variation) preclude a comprehensive analysis of pooled data. In order to be able to compare individual entries of such databases the batch effects need to be mitigated. Additionally, desired biological variation (referred to in this paper as " (experimental) design") between different independent experiments needs be conserved in any algorithm which aims to remove the batch effects.

Although a range of batch correction algorithms has previously been suggested [46, 28, 40, 8], only a small subset of these remains applicable in this large-scale setting. In particular, most previous algorithms cannot incorporate high-dimensional experimental design information. Here, we present a simple, yet effective adaptation of the popular empirical Bayes method [22] (ComBat) to account for a large amount highly

correlated biological covariates ("design features"). ComBat is based on ordinary linear regression and, therefore, will fail if the system is under determined.

We test our method on real-world microarray 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 [44] that thrives in a variety of environments and has been declared a critical priority pathogen for the development of new antimicrobial treatments [45]. A large range of studies have previously investigated the impact of culture conditions (e.g. growth media, temperature, oxygenation) on the gene expression profiles of PA. A comprehensive review of the perturbations caused by these different microenvironmental cues is missing as a consequence of the lack of harmonised data allowing for a direct comparison. The contributions of this paper can be summarised as follows.

- We extend the ComBat algorithm to handle many, highly correlated covariates, and (optionally) to perform feature selection.
- We address the issue of assessing the efficacy of the batch correction by investigating a range of suitable metrics.
- We present a large, harmonised data-set of PA expression profiles in response to different microenvironmental cues, and validate that biologically meaningful differences in transcription profiles are retained.

The paper is organised as follows. After reviewing relevant literature in Section 1.1 we introduce our *reComBat* algorithm in Section 2. In the second part of Section 2 we also introduce a large variety of evaluation metrics to quantify batch correction efficacy. In Section 3 we apply *reComBat* on a real-world data set of PA gene expression profiles. We conclude Section 3 by demonstrating, as a proof of concept, the biological validity of the harmonised data set. Our results are summarised and outlook on future research is given in Section 4.

1.1 Related Work

A large variety of batch correction methods has previously been suggested for both bulk and single cell sequencing data (see e.g. [28, 46, 50]). Here, we focus on batch correction methods for bulk data. In general, batch correction methods can be divided into the following categories:

- Normalisation to reference genes or samples
- Discretisation methods
- Location-scale adjustments
- Matrix factorisation
- Deep learning based

Algorithms, such as cross-platform normalization [42] or reference scaling [25], which employ references, are infeasible in the public data domain: "Reference" or "house keeping" genes may be impossible to define for some organisms, particularly microbes, eliminating the possibility to harness these as common ground for batch effect analysis and correction. Given a large (i.e. hundreds of independent experiments) public data set, it would be highly unlikely that overlapping samples or a common reference experiments are available for all of these.

Approaches that discretise expression data into categories (e.g., "expressed" vs. "not expressed") can be hard to implement rigorously without a relevant control. Furthermore, the information loss due to discretisation may affect the results of any advanced downstream analysis of the harmonised data.

Location-scale (LS) methods adjust the mean and/or variance of the genes. The simplest such methods use data standardisation [30] or batch mean-centring [43]. One of the most popular LS method is the empirical Bayes algorithm, ComBat [22]. Data standardisation generally only works if the batch effect is a simple mean/variance shift and it also does not account for additional confounders. 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. standardisation) or are unable to account for highly correlated design features (e.g. ComBat).

Matrix factorisation builds on decomposition approaches such as principal component analysis (PCA) or singular value decomposition (SVD) [2]. The aim is to identify and remove factors (singular vectors or

principal components) characterising the batch. While this approach 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 factorisation is to estimate unwanted variation via surrogate variables (SVA) [28]. Due to the fact that in our setting we assume that we know all sources of variation, we do not consider SVA.

Recently, nonlinear models, often based on neural autoencoders, have gained popularity (e.g. normAE [40], AD-AE [10], or scGEN [32]). Most models aim to find a batch-effect-free latent space representation of the data via adversarial training. While an advantage of these methods is their flexibility in being able to account for batches, but also desired biological variation, their major drawback is that the batch effect is only removed in a low-dimensional latent space and any downstream analysis is necessarily constrained. Further, in order to fully leverage the deep learning machinery, data sets with thousands of samples are needed.

2 Algorithm and Evaluation Metrics

In this section we introduce the mathematical and computational tools used in this paper. We start by defining our modification to the popular ComBat algorithm, *reComBat* and then we proceed to introducing a range of possible evaluation metrics to gauge the efficacy of data harmonisation.

2.1 Algorthim

ComBat: ComBat [22] is a well-established algorithm to perform batch correction of gene expression data. The raw data is, in essence, adjusted in a three-step process.

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

The ComBat algorithm has seen many refinements and applications to domains outside of microarray data (see e.g. [9, 35, 51]). However, most data sets have still been small and did not come with an extensive design matrix. When the design matrix becomes large (many covariates to consider) and sparse, unexpected issues can arise in step 1 of the algorithm. To illustrate the classic algorithm, we use the slightly modified ansatz of [49],

$$\boldsymbol{Y}_{ijk} = (\boldsymbol{X}\boldsymbol{\beta}^x)_{jk} + (\boldsymbol{C}\boldsymbol{\beta}^c)_{jk} + \boldsymbol{\alpha}_{jk} + \boldsymbol{\gamma}_{ik} + \boldsymbol{\delta}_{ik}\boldsymbol{\epsilon}_{ijk}, \tag{1}$$

where $\boldsymbol{Y}_{ijk}$ is the gene expression of the k^{th} gene in the j^{th} sample of the i^{th} 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 $\boldsymbol{\beta}^c$. The matrix $\boldsymbol{\alpha}$ is a matrix of intercepts, and $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ parameterise the *additive* and *multiplicative* batch effects. The tensor $\boldsymbol{\epsilon}$ is a three-dimensional tensor of standard Gaussian random variables. Note, that we implicitly encode batch- and sample-dependency by dropping the relevant indices, i.e. $\boldsymbol{\gamma}$ 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 (OLR) 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.

Once, the model is fitted, the data is standardised, 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 is adjusted. For details, please refer to the original publication [22].

reCombat: Using standard results for OLR, we know that the optimisation of (2) is convex, if and only if the matrix $\boldsymbol{A} = \tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T$ is non-singular. Therefore ComBat necessarily fails if $\boldsymbol{A}$ is rank-deficient, i.e. the system is underdetermined.

Based on 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 for batch correction of samples within the same experiment are limited, that is $m \ll n$. 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 taking all mutants of a particular wild type pathogen, which can number in the hundreds. 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 replicates (samples with identical experimental design) exist, then, even for large-scale integration, $\boldsymbol{A}$ may be rank deficient.

To mitigate this issue, we use standard approaches from linear regression theory and fit the elastic net model

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

$$\hat{\boldsymbol{\beta}}^x, \hat{\boldsymbol{\beta}}^c, \hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\beta}^x, \boldsymbol{\beta}^c, \boldsymbol{\alpha}}{\operatorname{argmin}} \left[\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|_2^2 + \lambda_1(\|\boldsymbol{\beta}^x\|_1 + \|\boldsymbol{\beta}^c\|_1) + \lambda_2(\|\boldsymbol{\beta}^x\|_2^2 + \|\boldsymbol{\beta}^c\|_2^2) \right], \tag{4}$$

where $\|\cdot\|_p$ denotes the ℓ_p norm, and λ_1 and λ_2 are the LASSO and ridge regularisation penalties. Due to this regularising modification of the algorithm we call our approach **re**gularised-ComBat, in short *reComBat*. Once the model is fitted the algorithm proceeds as usual.

2.2 Evaluating Batch Correction Efficacy

In the absence of a ground truth, quantification of the correctness of batch effect correction is difficult. Often, efficacy of a batch correction is judged by visual inspection, however, rigorous, quantitative evaluation of different aspects of the correction based on metrics are key.

This analysis is based on collections of samples with the same experimental design. Inspired by the graph theoretical notion of n-hop neighbourhoods [31], we group samples into so-called *Zero-Hops*. Each Zero-Hop defines a set of samples which share the exact same experimental design, a One-Hop would be a set of samples which differ from another in exactly one design condition, and so on. To this end, data harmonisation efficacy can be quantified in terms of Zero-Hop purity and batch impurity.

We implemented seven evaluation metrics quantifying the sample distance measures, cluster (im-)purity and batch/design classification performance on the obtained transcription profiles. Batch correction methods were compared with respect to the following metrics with statistical significance being evaluated by Mann-Whitney U test.

Sample distance- and neighbourhood-based metrics

Cross-distances: We defined a simple metric which calculates the median distance between all samples of a Zero-Hop and divides it by the median distance of all data points independent of batch or Zero-Hop. Correction methods resulting in smaller distances for Zero-Hops and large distances for batch indicate superior batch correction efficacy.

Distance Ratio Score: The cross-distance metric does not account for distances of individual samples but gives scores the median distance only. To complement this analysis we further assess the distance ratio score (DRS) [48]. The DRS quantifies the "closeness" of samples originating from the same condition versus the closeness of samples which should not. The DRS metric for a data-set of n samples is defined as

$$\mathrm{DRS}_{\log} = \frac{1}{n}\sum_{i=1}^{n} \log_2\left(\frac{d(\boldsymbol{Y}_i, \boldsymbol{Y}_{i,dt})}{d(\boldsymbol{Y}_i, \boldsymbol{Y}_{i,db/st})}\right), \tag{5}$$

where $d(\cdot,\cdot)$ is a distance metric, $\boldsymbol{Y}_i$ is the i^{th} sample and $\boldsymbol{Y}_{i,dt}$, $\boldsymbol{Y}_{i,db/st}$ are the closest samples from a different Zero-Hop (different type) and the same Zero-Hop (same type) but different batch respectively. A

good correction of batch effects hence results in an increase in $\mathrm{DRS}_{\log}$.
To account for the fact that $\boldsymbol{Y}_{i,db/st} = \inf$, i.e. no comparable inter-batch sample exists, we introduce the exponential-form DRS,

$$\mathrm{DRS}_{\exp} = \frac{1}{n}\sum_{i=1}^{n} \exp\left(\frac{d(\boldsymbol{Y}_i, \boldsymbol{Y}_{i,dt})}{d(\boldsymbol{Y}_i, \boldsymbol{Y}_{i,db/st})}\right) - 1. \tag{6}$$

Here, a dataset with no comparable inter-batch points has a DRS of 0 and the larger the metric the better the batch correction. Note, that modifications also exists when only comparing samples of the same type. We report the DRS for all Zero-Hops.

Shannon Entropy: Following [8], we calculated the local Shannon entropy with respect to batch and Zero-Hops within sets of N nearest neighbours to a sample k. A good batch correction algorithm strives for high batch entropy while maintaining a small Zero-Hop entropy. For each sample k the entropy S with respect to batch b is defined locally as

$$S_k = -\sum_{b=1}^{n_b} p_b \ln(p_b), \tag{7}$$

where p_b denotes the locally estimated probabilities of the different batches and n_b is the total number of batches. The Zero-Hop entropy is defined analogously. We chose the number of nearest neighbours for entropy calculation to be $N = 14$ corresponding to the median number of samples per Zero-Hop.

Cluster-based metrics

Minimum Separation Number: We define minimum separation number to be an integer quantifying the overlap of Zero-Hop clusters. Based on agglomerative clustering, initially, all samples occupy a single cluster. Then the number of clusters is increased in unit steps. The smallest number of clusters that assigns a Zero-Hop to at least two clusters, is defined as the minimum separation number for this Zero-Hop. We report mean values and standard deviations for the minimum number of separation steps required to separate all Zero-Hops.

Cluster purity: We first cluster the (corrected) expression data into n_{ZH} clusters, where n_{ZH} is the number of different Zero-Hops. The purity of each cluster c, $\mathrm{purity}(c)$, was calculated as the ratio of the number of samples of the dominant Zero-Hop in c, $n_{d,c}$, over the cluster size, n_c,

$$\mathrm{purity}(c) = \frac{n_{d,c}}{n_c}. \tag{8}$$

Gini impurity: The Gini impurity is a measure from decision tree learning [20, Section 9.2.3] and quantifies the probability of mislabelling a randomly chosen element of a cluster according to the label distribution of the cluster. Again, we create n_{ZH} clusters and use the Zero-Hop assignment as label. The output cluster impurities are given by assigning a label to each cluster chosen by a majority vote and calculating the fraction of majority labels in each cluster.

$$\mathrm{gini}(c) = 1 - \sum_{i=1}^{n_{\mathrm{ZH}}} p_i^2 \quad \text{with } p_i = \frac{n_{i,c}}{n_c}, \tag{9}$$

where $n_{i,c}$ is the number of samples of Zero-Hop i in cluster c.

Classifier-based metrics

Linear Discriminant Analysis: The Linear Discriminant Analysis (LDA) optimises hyperplanes to maximally separate data points according to their labels (here Zero-Hop) [20, Section 4.3]. We perform a stratified 10-fold cross validation and compare the LDA score (classification rate) on the held out test set. Mean values and standard deviations over the folds are reported.

Logistic-regression-based evaluation: This approach is inspired by the adversarial training of normAE [40]. In principle, any classifier should not be able to predict the batch from the gene expression profile of properly integrated data. Conversely, the prediction performance of the experimental design (Zero-Hop) should increase. To this end, we used two logistic regression (LR) classifiers, one for batch and one for Zero-Hops. Again, we perform a stratified 10-fold cross-validation and report the test set agglomerated balanced accuracy and F1-Scores. Mean values and standard deviations over the folds are reported.

3 Experiments

In this section, we apply ridge *reComBat* to a microarray data set of *Pseudomonas aeruginosa* (PA) gene expression data. We show quantitatively and qualitatively that *reComBat* is successful in removing the substantial batch effects while retaining biologically meaningful signal.

3.1 Data collection and preprocessing

Data was collected from the GEO database [7] (accessed October 2020). All entries on PA using the GPL84 Affymetrix GeneChip were considered. The GPL84 array comprises a total of 5900 probe sets including annotated genome of PA01 (5,568 genes) and other PA strains (117 genes), as well as 199 intergenic regions. In total, $n = 1260$ samples within 150 independent batches (GSE identifiers) were identified. The expression data were subjected to Robust Multi-array Averaging (RMA) using the `justRMA` function from the `affy` `R`-package. The relevant experimental design regarding culture conditions and PA strain was extracted manually and coarsened as outlined in Appendix A. For quality control (QC) we deleted all samples with incomplete design information and single-sample batches. Due to the large quantity of unique modifications, we dropped additional information regarding genetic alterations (mutations, plasmids, etc.) as part of the QC. As a final step of the preprocessing, we pruned all single-batch Zero-Hops.

3.2 Overview of the compiled data pool

Given the the imposed exclusion criteria, we analysed a total of 887 samples, structured within 39 Zero-Hops from 114 individual GSEs (i.e. batches) comprising 5 to 170 individual experiments each. An overview of the data set before and after quality control can be found in Table 1. Figure 1 A-C gives an overview of the included, uncorrected data coloured by batch, Zero-Hops, and PA strain. A detailed overview of all design categories contributing to the Zero-Hop definition is given in Appendix A Figure S1. As anticipated, we observe clustering by batch, rather than by Zero-Hops, indicating the presence of batch effects in addition to biologically meaningful variation.

Descriptor	**Original**	**After QC**
No. of samples	1260	887
No. of batches	150	114
No. of unique experimental designs	179	39

Table 1. Overview of the included data obtained from the GEO database before and after quality control to account for a minimum number of two samples per batch and a minimum number of two batches per unique experimental design.

3.3 Batch correction methods

Baselines: As baselines a broad array of applicable batch correction methods was used. In particular, we chose one location-scale, one matrix factorisation and one feature-based baseline. As mentioned in Subsection 1.1, normalisation to reference samples and deep learning-based methods were not applicable and discretisation would have resulted in too coarse features.

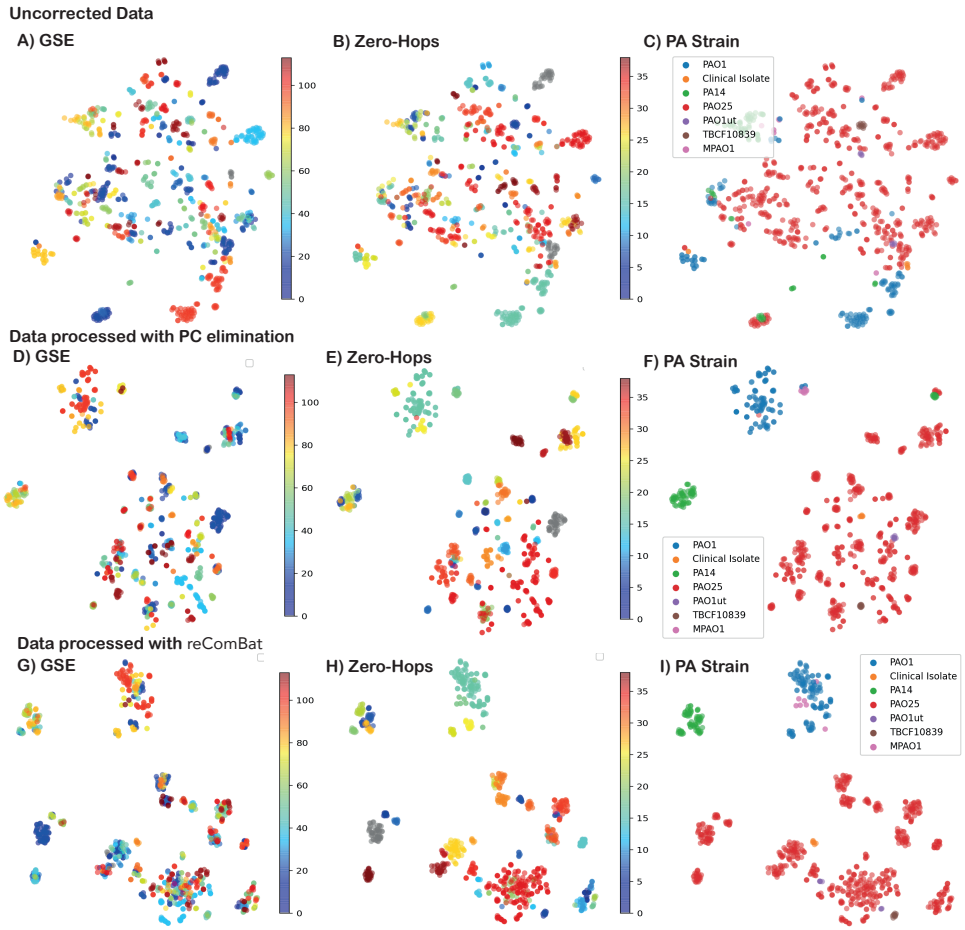

Fig. 1. tSNE plots of the uncorrected (A-C) and corrected (PC elimination: D-F, *reComBat* with $\lambda_1 = 0$, $\lambda_2 = 10^{-9}$: G-I) data, coloured by batches (A, D, G), Zero-Hops (B, E, H), and PA strain (C, F, I). For uncorrected data clustering is largely driven by the GSE, whereas the underlying culture conditions or microbial strains drive clusters on corrected data.

Standarization: One of the simplest location-scale methods is standardisation of each sample. For each sample we calculated the mean and standard deviation over all genes and z-scored the expression profiles accordingly. All subsequent methods took the standardised data as input.

Marker Gene elimination: This approach aims to eliminate genes which are highlighted as indicative genes of a batch. We first calculated marker genes between different batches and then eliminated the top eight marker genes of each batch pair from the data. This reduced the features to 5150 out of 5900 genes.

Principal component elimination: This method explicitly aims to bring Zero-Hops closer together. Samples within on Zero-Hop should have the same design and, therefore, the same gene expression profile. The first n principal components (PCs) of each Zero-Hop were calculated, with n being defined as 20% of the number of samples in the Zero-Hop, but a minimum of three. PCs accounting for more than 10% of the variance explained were subtracted from the data matrix.

reComBat: We used parametric priors for the empirical Bayes optimisation and we tested a variety of parameters for *reComBat*. In particular, we investigated pure LASSO ($\lambda_2 = 0$), pure ridge ($\lambda_1 = 0$), and the full elastic net regression. The range of regularisation 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 the smaller values of λ_1 yielded numerical instabilities. We report the best performing configuration only.

3.4 Evaluation of the harmonised dataset

In addition to a quantitative comparison based on the metrics described in section 2.2, we provide a notion of whether biologically meaningful expression profiles are retained following batch correction. Biological validation of the output is crucial as batch correction methods can be prone to overfitting [54]. As representative examples we analysed data subsets differing either in terms of oxygenation status, culture medium richness, growth phase, or clinical vs. laboratory PA strains. We find the marker genes for the Zero-Hops and evaluate the top 50 genes driving the differences between selected pairwise comparisons. We then identify the relevant underlying biological pathways driving differences.

Performance assessment from our metrics: We compared our batch correction baselines to the best performing *reComBat* models. In order to quantify the correction success, we calculated the evaluation metrics of Subsection 2.2. Our results are summarised in Figure 2.

Data standardisation and marker gene elimination only had a minor, insignificant (all p-values > 0.05) effect when compared to the raw data. PC elimination led to markedly improved evaluation metrics and, hence, provided reasonable batch correction (see also Appendix B). However, despite significant (all p-values < 0.001) improvement of PC elimination compared to raw data large variations were observed for different Zero-Hops throughout all distance-based metrics for this correction method.

reComBat batch correction success depended on the regularisation type and strength imposed. Best results for all evaluated metrics were obtained for ridge regularisation ($\lambda_1 = 0$) with $\lambda_2 \leq 0.001$. Below this threshold any observed improvements were minor. One possible explanation is due to the fact that, as $\lambda_2 \to 0$, the ridge regression approximates the OLR, i.e. an unregularised, solution. Whereas both LASSO and Elastic Net *reComBat* were inferior to PC elimination, ridge-*reComBat* outperformed PC elimination in several of the assessed metrics with small variation across different Zero-Hops or clusters (see Figure 2). We observe that stronger, particular LASSO, regularisation achieves superior batch heterogeneity (i.e. low classification performance) but this comes at the cost of decrease in Zero-Hop performance. As such, ridge regularisation was deemed most suitable for this experiment. Notice the LASSO-*reComBat* performs implicit feature selection due the ℓ_1 regularisation. This could hint to the fact that more balanced feature weighting (as provided by ridge-*reComBat*) is beneficial.

These positive results for ridge *reComBat* can also be observed by visual inspection of tSNE plots (see Figure 1D-F). Separated clusters based on the underlying culture conditions and microbial strains are achieved. Detailed tSNE plots coloured for all design matrix elements for all baselines and varieties of *reComBat* addressed by evaluation metrics are provided in Appendix B, Figures S3 -S7.

In terms of the gauging of the metrics themselves for the ability to detect batch effects, we conclude that classifier-based metrics are far superior to any other approaches. Shannon entropy can detect a larger spread in batch vs. 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 neighbours. Likewise, the median pairwise distance and DRS metrics show some ability to detect batch correction, but due to the strong dependency on the Zero-Hop the spread in values may also be large. The minimum separation clustering clearly shows when a batch correction can be considered effective. However, due to repeated clustering (up to number-of-batches many clusterings), calculation of minimum separation number is computationally far more expensive than distance based metrics. A good mid-point metrics between classifier-based evaluation and cluster-based evaluation seem to be the cluster-purity measures, which show good resolution and manageable dependency on the Zero-Hop. Of course, any of these metrics could be accumulated to a single number via averaging over the Zero-Hops.

Characterisation of the harmonised dataset In order to preclude over-correcting the data it is essential to demonstrate that biologically meaningful expression profiles are retained following batch correction. Indeed we were able to show that pathways previously known to be important in the relevant culture conditions were identified (see Supplement for a complete lists of the extracted genes for each comparison). For instance, when comparing standard growth conditions (PA grown in liquid LB, while shaking in aerobic conditions at 37°C, in exponential growth phase) to hypoxia conditions, we find that genes involved in aerotaxis (PA1561) [21], Fe-S cluster biogenesis (PA4615) [39] and iron acquisition (PA2391, PA2399, PA2407, PA2948, PA3404-3405, PA3407-3408) [15, 17, 18, 34, 36] are major drivers of differences. When comparing

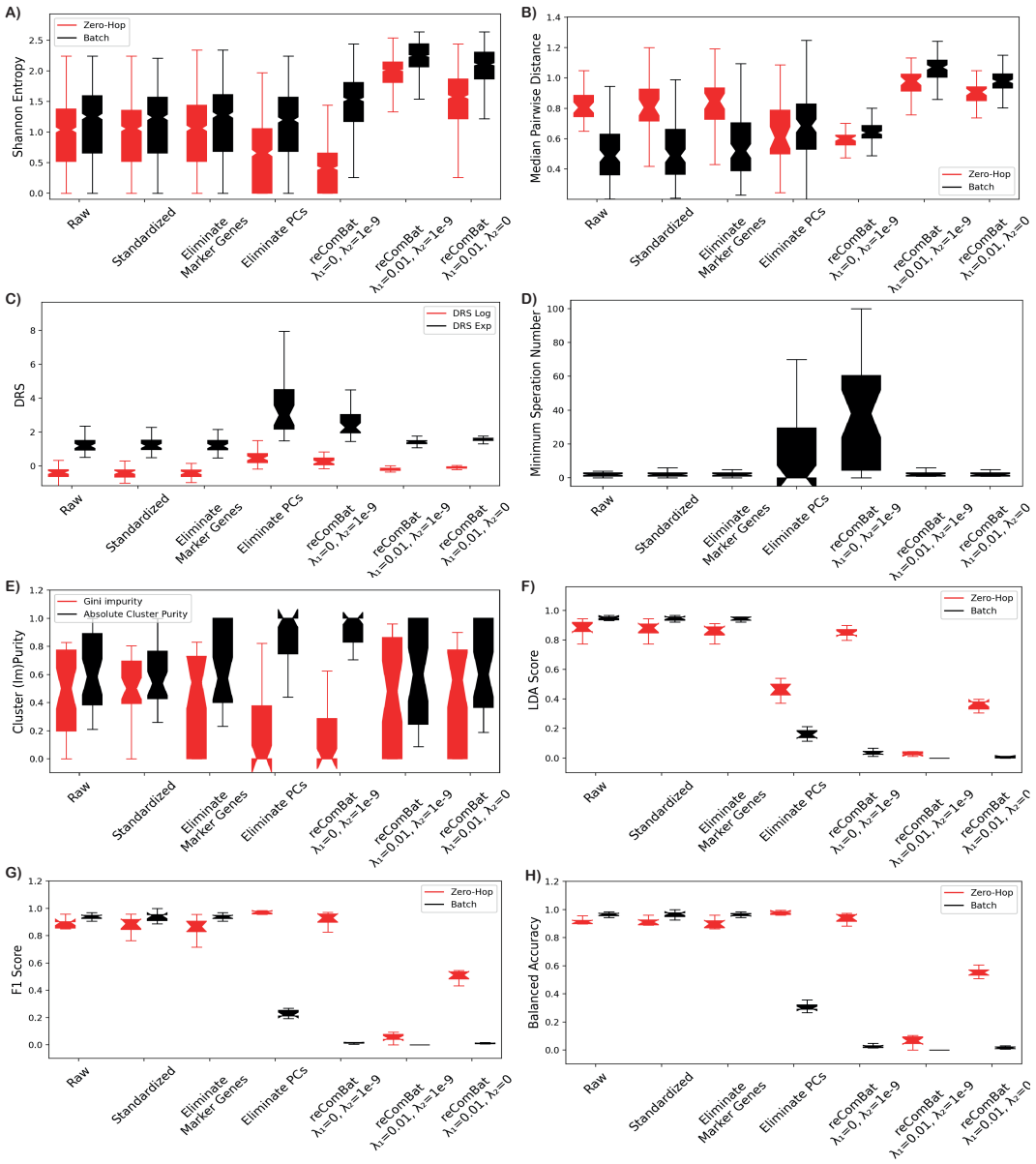

Fig. 2. Evaluation metrics scoring the impact of batch effects by evaluating the variety of different batches and/or Zero-Hops of a specific sample. Box plots represent the lower and upper quartiles (box) together with the median (central dents) and full range (whiskers) over all samples, clusters, or Zero-Hops depending on the relevant metric. LDA scores and LR classification performance are given reported over ten cross validation folds (F, G, H).

cultures in exponential to stationary phase under hypoxia conditions, genes involved in pyoverdin (PA2412-2413, PA2426) [11, 47] and pyochelin (PA4221-4226, PA4228-4231) [4, 14, 38] biosynthesis and transport, iron starvation (PA0197, PA2384, PA3407, PA4468-PA4469) [1, 19, 52, 53] and quorum sensing (PA2512-2514) [26] were relevant. Finally, for a comparison between exponential growth of the laboratory strain PAO1 vs. clinical isolates in rich media (37°C, aerobic) we find cup genes (PA4081-PA4084, PA0994) that are involved in motility and attachment and with this in biofilm formation [41]. This indicates a difference in attachment between those strains that might be coming form the environment the strains have adapted to grow in (laboratory vs. 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 harmonised data set hence serves for hypothesis generation motivating further (experimental) validation. By mining the harmonised data set, we can also perform comparisons that have, to the best of our knowledge, never been directly performed before. For instance, when we compare growth in LB with growth in media that have fewer nutrients than standard LB, we find that several nutrient uptake pathways are of particular importance. In particular, genes involved in phosphorous uptake and stress (PA0676, PA0680, PA0683, PA0686, PA0690-0693, PA0842, PA3372, PA3375-3376, PA3378-3379, PA3381-3384, PA4350) [5, 6, 13, 23, 29, 37] and metal uptake (PA2399, PA3407) [1, 34] are deferentially regulated. As such, we were able to demonstrate that biologically meaningful information was retained in the harmonised data.

4 Discussion

Public databases play an increasingly important role for data-driven meta-analysis in the field of computational biology. Despite great efforts of harmonising data collection, considerable, yet unavoidable, biological/technical variation may mask true signal if data are pooled from several sources. Aiming at drawing generalisable conclusions from agglomerated data sets, it is essential to correct such batch effects in a setting where overlapping samples, or standardised controls, are unavailable. Often large number of batches coincides with desired biological variation, which renders a range of standard batch correction algorithms inapplicable. Here, we presented a simple, yet effective, means of mitigating highly correlated experimental conditions through regularisation and compared various elastic net regularisation strengths for this purpose.

reComBat outperformed standard approaches for large scale batch correction, including data standardisation, principal component and marker gene elimination with respect to the design subsets under study. 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 for the overall performance of our algorithm. Moreover, we demonstrate adequate batch correction performance while retaining desired biological signals, as is confirmed by manual inspection of the harmonised data.

A limitation of the biological evaluation is the fact that performing marker gene analysis between the Zero-Hops is done on the corrected data. This procedure might implicitly bias the $p - values$ obtained for the marker genes as any batch correction method reduces the degrees of freedom of the data - as such we specifically included all of the top 50 scoring genes in our literature search, irrespective of significance levels at this point. Traditionally, differential gene expression analysis is performed on uncorrected data with relevant confounders given as an input. However, given the complexity of the task it may be essential to correct batch effects first, before performing any marker gene analysis. Despite our simplified evaluation, it was possible to show that biologically meaningful target pathways could be identified. Yet, our evaluation was limited to literature-based validation of the identified driver pathways. Experimental validation of the proposed findings is key in confirming information on the underlying biological mechanisms. Here, the identified differences in gene expression may serve mostly the purpose of hypothesis generation.

Another possible use of *reComBat* is its application to RNA sequencing and other types of omics data. Although similar numbers of samples are publicly available for bulk RNA sequencing of PA under varying culture conditions, additional issues, including annotation standardisation between strains leads to great preprocessing efforts which we are currently further investigating. We also would like to stress the time intensive manual work involved in the collection of the experimental design to provide an optimal data annotation.

In this work, we deliberately decided against the application of deep or nonlinear models as in the setting of bulk sequencing data the number of samples may be insufficient for model training (ranging in the order of hundreds, rather than thousands). Hence, increasing model complexity may result in overfitting of the batch correction. We would like to stress, that in case of e.g. single cell RNA sequencing experiments, or larger data sets, the situation may indeed be favourable for nonlinear approaches. Investigating non-linear interaction in experimental design constitutes part of future research.

From an application perspective, however, we showed how a straightforward adaption of the popular ComBat algorithm can drastically increase its usability. ComBat benefits from low computational cost and rigorous underlying theory and it is easy to apply in practice. By further publishing *reComBat* as a python

package[4] our method is readily available to the community. We also make the harmonised data set with all its metadata available to the wider research community.[5] This will allow researchers to gain novel insights into the behaviour of the incredibly adaptable PA - a key for developing new drugs against this increasingly resistant pathogen.

Data availability

We make *reComBat* available as python package (https://github.com/BorgwardtLab/reComBat.git) and have also published the code, harmonized and uncorreced data in the following repository: https://github.com/BorgwardtLab/batchCorrectionPublicData.git

Acknowledgements

This project was supported by the National Center of Competence in Research AntiResist funded by the Swiss National Science Foundation (grant number 51NF40_180541) and funded in part by the Alfried Krupp Prize for Young University Teachers of the Alfried Krupp von Bohlen und Halbach-Stiftung (K.B.).

References

1. Alontaga, A.Y., Rodriguez, J.C., Schönbrunn, E., Becker, A., Funke, T., Yukl, E.T., Hayashi, T., Stobaugh, J., Moënne-Loccoz, P., Rivera, M.: Structural characterization of the hemophore HasAp from Pseudomonas aeruginosa: NMR spectroscopy reveals protein-protein interactions between Holo-HasAp and hemoglobin. Biochemistry **48**(1), 96–109 (jan 2009). https://doi.org/10.1021/bi801860g
2. Alter, O., Brown, P.O., Botstein, D.: Singular value decomposition for genome-wide expression data processing and modeling. Proceedings of the National Academy of Sciences of the United States of America **97**(18), 10101–6 (aug 2000). https://doi.org/10.1073/pnas.97.18.10101
3. Andersson, D.I., Balaban, N.Q., Baquero, F., Courvalin, P., Glaser, P., Gophna, U., Kishony, R., Molin, S., Tnjum, T.: Antibiotic resistance: turning evolutionary principles into clinical reality. FEMS Microbiology Reviews **44**(2), 171–188 (01 2020). https://doi.org/10.1093/femsre/fuaa001
4. Ankenbauer, R.G., Quan, H.N.: FptA, the Fe(III)-pyochelin receptor of Pseudomonas aeruginosa: a phenolate siderophore receptor homologous to hydroxamate siderophore receptors. Journal of bacteriology **176**(2), 307–19 (jan 1994). https://doi.org/10.1128/jb.176.2.307-319.1994
5. Bains, M., Fernández, L., Hancock, R.E.W.: Phosphate starvation promotes swarming motility and cytotoxicity of Pseudomonas aeruginosa. Applied and environmental microbiology **78**(18), 6762–8 (sep 2012). https://doi.org/10.1128/AEM.01015-12
6. Ball, G., Durand, E., Lazdunski, A., Filloux, A.: A novel type II secretion system in Pseudomonas aeruginosa. Molecular microbiology **43**(2), 475–85 (jan 2002). https://doi.org/10.1046/j.1365-2958.2002.02759.x
7. Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C.L., Serova, N., Davis, S., Soboleva, A.: NCBI GEO: archive for functional genomics data sets–update. Nucleic acids research **41**(Database issue), D991–5 (jan 2013). https://doi.org/10.1093/nar/gks1193
8. Chazarra-Gil, R., van Dongen, S., Kiselev, V.Y., Hemberg, M.: Flexible comparison of batch correction methods for single-cell rna-seq using batchbench. Nucleic acids research **49**(7), e42–e42 (2021)
9. Čuklina, J., Lee, C.H., Williams, E.G., Sajic, T., Collins, B.C., Rodríguez Martínez, M., Sharma, V.S., Wendt, F., Goetze, S., Keele, G.R., Wollscheid, B., Aebersold, R., Pedrioli, P.G.A.: Diagnostics and correction of batch effects in largescale proteomic studies: a tutorial. Molecular Systems Biology **17**(8) (aug 2021). https://doi.org/10.15252/msb.202110240
10. Dincer, A.B., Janizek, J.D., Lee, S.I.: Adversarial deconfounding autoencoder for learning robust gene expression embeddings. Bioinformatics **36**(Supplement_2), i573–i582 (2020)
11. Drake, E.J., Cao, J., Qu, J., Shah, M.B., Straubinger, R.M., Gulick, A.M.: The 1.8 Å Crystal Structure of PA2412, an MbtH-like Protein from the Pyoverdine Cluster of Pseudomonas aeruginosa. Journal of Biological Chemistry **282**(28), 20425–20434 (jul 2007). https://doi.org/10.1074/jbc.M611833200

[4] https://github.com/BorgwardtLab/reComBat.git
[5] https://github.com/BorgwardtLab/batchCorrectionPublicData.git

12. Edgar, R., Domrachev, M., Lash, A.E.: Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research **30**(1), 207–10 (jan 2002). https://doi.org/10.1093/nar/30.1.207
13. Faure, L.M., Garvis, S., de Bentzmann, S., Bigot, S.: Characterization of a novel two-partner secretion system implicated in the virulence of Pseudomonas aeruginosa. Microbiology (Reading, England) **160**(Pt 9), 1940–1952 (sep 2014). https://doi.org/10.1099/mic.0.079616-0
14. Gaille, C., Reimmann, C., Haas, D.: Isochorismate synthase (PchA), the first and rate-limiting enzyme in salicylate biosynthesis of Pseudomonas aeruginosa. The Journal of biological chemistry **278**(19), 16893–8 (may 2003). https://doi.org/10.1074/jbc.M212324200
15. Gao, L., Guo, Z., Wang, Y., Wang, Y., Wang, K., Li, B., Shen, L.: The Two-Operon-Coded ABC Transporter Complex FpvWXYZCDEF is Required for Pseudomonas aeruginosa Growth and Virulence Under Iron-Limiting Conditions. The Journal of membrane biology **251**(1), 91–104 (2018). https://doi.org/10.1007/s00232-017-9979-5
16. Gil-Gil, T., Ochoa-Snchez, L.E., Baquero, F., Martnez, J.L.: Antibiotic resistance: Time of synthesis in a post-genomic age. Computational and Structural Biotechnology Journal **19**, 3110–3124 (2021). https://doi.org/https://doi.org/10.1016/j.csbj.2021.05.034
17. Glanville, D.G., Mullineaux-Sanders, C., Corcoran, C.J., Burger, B.T., Imam, S., Donohue, T.J., Ulijasz, A.T.: A High-Throughput Method for Identifying Novel Genes That Influence Metabolic Pathways Reveals New Iron and Heme Regulation in Pseudomonas aeruginosa. mSystems **6**(1) (feb 2021). https://doi.org/10.1128/mSystems.00933-20
18. Hannauer, M., Braud, A., Hoegy, F., Ronot, P., Boos, A., Schalk, I.J.: The PvdRT-OpmQ efflux pump controls the metal selectivity of the iron uptake pathway mediated by the siderophore pyoverdine in Pseudomonas aeruginosa. Environmental microbiology **14**(7), 1696–708 (jul 2012). https://doi.org/10.1111/j.1462-2920.2011.02674.x
19. Hassett, D.J., Howell, M.L., Sokol, P.A., Vasil, M.L., Dean, G.E.: 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. Journal of bacteriology **179**(5), 1442–51 (mar 1997). https://doi.org/10.1128/jb.179.5.1442-1451.1997
20. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer Series in Statistics, Springer New York Inc., New York, NY, USA (2001)
21. Hong, C.S., Shitashiro, M., Kuroda, A., Ikeda, T., Takiguchi, N., Ohtake, H., Kato, J.: Chemotaxis proteins and transducers for aerotaxis in Pseudomonas aeruginosa. FEMS microbiology letters **231**(2), 247–52 (feb 2004). https://doi.org/10.1016/S0378-1097(04)00009-6
22. Johnson, W.E., Li, C., Rabinovic, A.: Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics **8**(1), 118–127 (04 2006). https://doi.org/10.1093/biostatistics/kxj037
23. Jones, R.A., Shropshire, H., Zhao, C., Murphy, A., Lidbury, I., Wei, T., Scanlan, D.J., Chen, Y.: Phosphorus stress induces the synthesis of novel glycolipids in Pseudomonas aeruginosa that confer protection against a last-resort antibiotic. The ISME Journal **15**(11), 3303–3314 (nov 2021). https://doi.org/10.1038/s41396-021-01008-7
24. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al.: Highly accurate protein structure prediction with alphafold. Nature **596**(7873), 583–589 (2021)
25. Kim, K.Y., Kim, S.H., Ki, D.H., Jeong, J., Jeong, H.J., Jeung, H.C., Chung, H.C., Rha, S.Y.: An attempt for combining microarray data sets by adjusting gene expressions. Cancer research and treatment **39**(2), 74–81 (jun 2007). https://doi.org/10.4143/crt.2007.39.2.74
26. Kim, S.K., Im, S.J., Yeom, D.H., Lee, J.H.: AntR-mediated bidirectional activation of antA and antR, anthranilate degradative genes in Pseudomonas aeruginosa. Gene **505**(1), 146–52 (aug 2012). https://doi.org/10.1016/j.gene.2012.05.004
27. Kourou, K., Exarchos, K.P., Papaloukas, C., Sakaloglou, P., Exarchos, T., Fotiadis, D.I.: Applied machine learning in cancer research: A systematic review for patient diagnosis, classification and prognosis. Computational and Structural Biotechnology Journal **19**, 5546–5555 (2021). https://doi.org/https://doi.org/10.1016/j.csbj.2021.10.006
28. Lazar, C., Meganck, S., Taminau, J., Steenhoff, D., Coletta, A., Molter, C., Weiss-Sols, D.Y., Duque, R., Bersini, H., Now, A.: Batch effect removal methods for microarray gene expression data integration: a survey. Briefings in Bioinformatics **14**(4), 469–490 (07 2012). https://doi.org/10.1093/bib/bbs037
29. Lewenza, S., Falsafi, R., Bains, M., Rohs, P., Stupak, J., Sprott, G.D., Hancock, R.E.W.: 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 microbiology letters **320**(2), 95–102 (jul 2011). https://doi.org/10.1111/j.1574-6968.2011.02295.x
30. Li, C., Wong, W.H.: Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proceedings of the National Academy of Sciences of the United States of America **98**(1), 31–6 (jan 2001). https://doi.org/10.1073/pnas.011404098
31. Liu, W., Li, Z.: An efficient parallel algorithm of n-hop neighborhoods on graphs in distributed environment. Frontiers of Computer Science **13**(6), 1309–1325 (2019)

32. Lotfollahi, M., Wolf, F.A., Theis, F.J.: scGen predicts single-cell perturbation responses. Nature methods **16**(8), 715–721 (2019). https://doi.org/10.1038/s41592-019-0494-8
33. Malod-Dognin, N., Petschnigg, J., Windels, S.F.L., Povh, J., Hemingway, H., Ketteler, R., Pržulj, N.: Towards a data-integrated cell. Nature Communications **10**(1), 805 (dec 2019). https://doi.org/10.1038/s41467-019-08797-8
34. Merriman, T.R., Merriman, M.E., Lamont, I.L.: Nucleotide sequence of pvdD, a pyoverdine biosynthetic gene from Pseudomonas aeruginosa: PvdD has similarity to peptide synthetases. Journal of bacteriology **177**(1), 252–8 (jan 1995). https://doi.org/10.1128/jb.177.1.252-258.1995
35. Müller, C., Schillert, A., Röthemeier, C., Trégouët, D.A., Proust, C., Binder, H., Pfeiffer, N., Beutel, M., Lackner, K.J., Schnabel, R.B., Tiret, L., Wild, P.S., Blankenberg, S., Zeller, T., Ziegler, A.: Removing Batch Effects from Longitudinal Gene Expression - Quantile Normalization Plus ComBat as Best Approach for Microarray Transcriptome Data. PLOS ONE **11**(6), e0156594 (jun 2016). https://doi.org/10.1371/journal.pone.0156594
36. Ochsner, U.A., Johnson, Z., Vasil, M.L.: Genetics and regulation of two distinct haem-uptake systems, phu and has, in Pseudomonas aeruginosa. Microbiology (Reading, England) **146 (Pt 1**, 185–198 (jan 2000). https://doi.org/10.1099/00221287-146-1-185
37. Quesada, J.M., Otero-Asman, J.R., Bastiaansen, K.C., Civantos, C., Llamas, M.A.: The Activity of the Pseudomonas aeruginosa Virulence Regulator σVreI Is Modulated by the Anti-σ Factor VreR and the Transcription Factor PhoB. Frontiers in Microbiology **7** (aug 2016). https://doi.org/10.3389/fmicb.2016.01159
38. Reimmann, C., Patel, H.M., Serino, L., Barone, M., Walsh, C.T., Haas, D.: Essential PchG-dependent reduction in pyochelin biosynthesis of Pseudomonas aeruginosa. Journal of bacteriology **183**(3), 813–20 (feb 2001). https://doi.org/10.1128/JB.183.3.813-820.2001
39. Romsang, A., Duang-nkern, J., Wirathorn, W., Vattanaviboon, P., Mongkolsuk, S.: Pseudomonas aeruginosa IscR-Regulated Ferredoxin NADP(+) Reductase Gene (fprB) Functions in Iron-Sulfur Cluster Biogenesis and Multiple Stress Response. PLOS ONE **10**(7), e0134374 (jul 2015). https://doi.org/10.1371/journal.pone.0134374
40. Rong, Z., Tan, Q., Cao, L., Zhang, L., Deng, K., Huang, Y., Zhu, Z.J., Li, Z., Li, K.: NormAE: Deep Adversarial Learning Model to Remove Batch Effects in Liquid Chromatography Mass Spectrometry-Based Metabolomics Data. Analytical chemistry **92**(7), 5082–5090 (2020). https://doi.org/10.1021/acs.analchem.9b05460
41. Ruer, S., Stender, S., Filloux, A., de Bentzmann, S.: Assembly of fimbrial structures in pseudomonas aeruginosa: Functionality and specificity of chaperone-usher machineries. Journal of Bacteriology **189**(9), 3547–3555 (2007). https://doi.org/10.1128/JB.00093-07
42. Shabalin, A.A., Tjelmeland, H., Fan, C., Perou, C.M., Nobel, A.B.: Merging two gene-expression studies via cross-platform normalization. Bioinformatics (Oxford, England) **24**(9), 1154–60 (may 2008). https://doi.org/10.1093/bioinformatics/btn083
43. Sims, A.H., Smethurst, G.J., Hey, Y., Okoniewski, M.J., Pepper, S.D., Howell, A., Miller, C.J., Clarke, R.B.: The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets improving meta-analysis and prediction of prognosis. BMC Medical Genomics **1**(1), 42 (dec 2008). https://doi.org/10.1186/1755-8794-1-42
44. Stover, C.K., Pham, X.Q., Erwin, A.L., Mizoguchi, S.D., Warrener, P., Hickey, M.J., Brinkman, F.L., Hufnagle, W.O., Kowalik, D.J., Lagrou, M., Garber, R.L., Goltry, L., Tolentino, E., Westbrock-Wadman, S., Yuan, Y., Brody, L.L., Coulter, S.N., Folger, K.R., Kas, A., Larbig, K., Lim, R., Smith, K., Spencer, D., Wong, G.K.S., Wu, Z., Paulsen, I.T., Reizer, J., Saier, M.H., Hancock, R.E.W., Lory, S., Olson, M.V.: Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen. Nature **406**(6799), 959–964 (aug 2000). https://doi.org/10.1038/35023079
45. Tacconelli, E., Carrara, E., Savoldi, A., Harbarth, S., Mendelson, M., Monnet, D.L., Pulcini, C., Kahlmeter, G., Kluytmans, J., Carmeli, Y., Ouellette, M., Outterson, K., Patel, J., Cavaleri, M., Cox, E.M., Houchens, C.R., Grayson, M.L., Hansen, P., Singh, N., Theuretzbacher, U., Magrini, N., WHO Pathogens Priority List Working Group: Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. The Lancet. Infectious diseases **18**(3), 318–327 (2018). https://doi.org/10.1016/S1473-3099(17)30753-3
46. Tran, H.T.N., Ang, K.S., Chevrier, M., Zhang, X., Lee, N.Y.S., Goh, M., Chen, J.: A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biology **21**(1), 12 (dec 2020). https://doi.org/10.1186/s13059-019-1850-9
47. Vandenende, C.S., Vlasschaert, M., Seah, S.Y.K.: Functional characterization of an aminotransferase required for pyoverdine siderophore biosynthesis in Pseudomonas aeruginosa PAO1. Journal of bacteriology **186**(17), 5596–602 (sep 2004). https://doi.org/10.1128/JB.186.17.5596-5602.2004
48. Varma, S.: Blind estimation and correction of microarray batch effect. PLOS ONE **15**(4), e0231446 (apr 2020). https://doi.org/10.1371/journal.pone.0231446
49. Wachinger, C., Rieckmann, A., Pölsterl, S., Initiative, A.D.N., et al.: Detect and correct bias in multi-site neuroimaging datasets. Medical Image Analysis **67**, 101879 (2021)

50. Yu, X., Abbas-Aghababazadeh, F., Chen, Y.A., Fridley, B.L.: Statistical and Bioinformatics Analysis of Data from Bulk and Single-Cell RNA Sequencing Experiments, pp. 143–175. Springer US, New York, NY (2021). https://doi.org/10.1007/978-1-0716-0849-4_9
51. Zhang, Y., Parmigiani, G., Johnson, W.E.: ComBat-seq: batch effect adjustment for RNA-seq count data. NAR genomics and bioinformatics **2**(3), lqaa078 (sep 2020). https://doi.org/10.1093/nargab/lqaa078
52. Zhao, Q., Poole, K.: A second tonB gene in Pseudomonas aeruginosa is linked to the exbB and exbD genes. FEMS Microbiology Letters **184**(1), 127–132 (mar 2000). https://doi.org/10.1111/j.1574-6968.2000.tb09002.x
53. Zheng, P., Sun, J., Geffers, R., Zeng, A.P.: Functional characterization of the gene PA2384 in large-scale gene regulation in response to iron starvation in Pseudomonas aeruginosa. Journal of Biotechnology **132**(4), 342–352 (dec 2007). https://doi.org/10.1016/j.jbiotec.2007.08.013
54. Zindler, T., Frieling, H., Neyazi, A., Bleich, S., Friedel, E.: Simulating ComBat: how batch correction can lead to the systematic introduction of false positive results in DNA methylation microarray studies. BMC Bioinformatics **21**(1), 271 (dec 2020). https://doi.org/10.1186/s12859-020-03559-6