Batch Effect Correction in a Confounded Scenario: a Case Study on Gene Expression of Chornobyl Tree Frogs

Elen Goujon, Olivier Armant, Clément Car, Jean-Marc Bonzom, Arthur Tenenhaus, Imène Garali

Batch Effect Correction in a Confounded Scenario: a Case Study on Gene Expression of Chornobyl Tree Frogs

Elen Goujon[1,3][0000−0002−4237−5754], Olivier Armant[2][0000−0001−7101−9209], Clément Car[2][0000−0002−2729−2950], Jean-Marc Bonzom[2][0000−0002−6526−5769], Arthur Tenenhaus[3][0000−0003−3459−1518], and Imène Garali[1][0000−0001−5779−5199]

[1] French Institute for Radiation Protection and Nuclear Safety (IRSN), PSE-SANTE/SESANE/LRTOX, 92260 Fontenay-aux-Roses, France
[2] IRSN, PSE-ENV/SERPEN/LECO, 13115 Saint-Paul-Lez-Durance, France
[3] Laboratoire des Signaux et Systèmes, Université Paris-Saclay, CNRS, CentraleSupélec, 91190 Gif-sur-Yvette, France
imene.garalizineddine@irsn.fr

Abstract. When large omics datasets present unwanted latent variability, a critical analysis step is to control these so-called batch effects properly. However, most batch effects-correction algorithms (BECAs) face limitations when the source of unwanted variation and the variable of interest are confounded. In this paper, we use RNA-seq data to study the effects of radiation contamination on tree frogs (*Hyla orientalis*) collected in the Chornobyl Exclusion Zone. We identify the site of collection of the frogs as a confounding factor in the transcriptomics analysis. We present our strategy to correct this confounding effect using the following BECAs: ComBat-seq, linear residualization, and Surrogate Variable Analysis. We show that the severe confounding between the site and radiocontamination level makes the correction step challenging. Instead, we investigate the site-to-site variability and successfully deconvolute the batch variable from the radiation level by adjusting for the population genetic structure. Our strategy allowed us to reveal the effects of low-dose radiation on the gene expression of Chornobyl tree frogs and appropriately preprocess the RNA-seq dataset for future multimodal integrative analyses.

Keywords: Confounding factors · Batch effect-correction algorithms · Transcriptomics · Chornobyl tree frogs · Low-dose radiation

1 Introduction

In the current age of research in biology, high-throughput data such as omics measurements represent an immense wealth of information, assuming that they are processed appropriately. The analysis of such datasets often aims at identifying differential responses between classes of study – for example, groups of treatment, disease, or levels of exposure to a pollutant. However, unwanted variation can arise in datasets, stemming from technical or biological origins, and be

detrimental to the analysis by leading to misguided interpretations [12, 13, 21]. The term "batch effects" describes this underlying variability, and their handling has become a crucial step in data analysis.

Batch effects can be tackled by one of two strategies. Either include the batch in the model as a covariate in a one-step approach or correct for the batch in a pre-processing step using batch-effect correction algorithms (BECAs) before any downstream analysis [24, 29]. In RNA-Seq data analysis, various differential expression methods allow the integration of batch effects and covariates in the model design [9, 25, 30]. While one-step methods can be preferred because of their simplicity, it is not always possible to integrate covariates, depending on the model type. Two-step batch adjustment allows more flexibility, and the myriad of bioinformatics tools dedicated to batch-effect correction should allow one to choose a context-appropriate BECA [41].

Still, difficulties emerge in scenarios where the batch and the study class or variable of interest have a simultaneous influence on the data. If batch and class are confounded, then adjusting the dataset to differentiate the effects of one from the other becomes more challenging [12, 29, 39, 41]. BECA comparison studies have produced conflicted results in the case of batch-class design imbalance, with tools such as ComBat [16], Surrogate Variable Analysis (SVA) [22], batch mean-centering (BMC) [32], and ratio-based methods performing differently from one example to another [39, 41].

In this study, we use a published dataset on the effects of chronic exposure to low-dose radiation on wildlife in the aftermath of the nuclear accident in Chornobyl [3, 8]. In particular, 87 tree frogs from species *Hyla orientalis* were sampled in 8 sites inside and outside of the Chornobyl Exclusion Zone (CEZ) in 2018 [8]. Previous research efforts have shown the presence of recent evolutionary processes at play in the CEZ [7]. Also, altered transcriptomic and metabolic pathways were identified, as well as changes in physiological traits for the frogs living in the most contaminated sites [8].

Different data modalities have been obtained, and we plan to conduct a multi-omics data integration study using transcriptomics, proteomics, and genomics data. In particular, for RNA-Seq expression data, we have found that this dataset is impacted by the site of collection of the frogs in a batch effect-like manner. Given the distribution of the geographical sites in the Chornobyl area, the batch is highly correlated with the level of exposure (µGy/h), which is our variable of interest. Before integrating data in a multi-omics study, we want to identify an appropriate approach for handling this batch effect.

In this paper, we present three different strategies to correct for the effect of the site. First, we adjusted the count matrix for the site effect using classical techniques such as ComBat-seq [40] and linear residualization [11], with and without integrating exposure levels in the correction step. In a second strategy, we dissected the site variable to decouple the site effects from the radiation dose rate. We provide a new approach by integrating another data source: the genetic distance between individuals, which partly explains the site variability, is treated as a batch effect and corrected. Lastly, we use an exploratory approach

to estimate the directions of systematic expression variation not explained by radiation exposure, thanks to SVA [22].

We structure the remaining parts as follows. In Section 2, we present the data handled in this study. Section 3 introduces the methods used for batch-effect correction, batch-effect assessment, and performance evaluation. The results of the different correction strategies are shown in Section 4, along with statistical and biological comparisons. We further discuss our results in Section 5 and conclude in Section 6.

2 Tree Frog Sampling and Dataset Description

The 87 individuals studied here are male Eastern tree frogs (*Hyla orientalis*) collected in Ukraine in 2018 across a gradient of radioactive contamination. Population sampling was done in 8 different wetlands in the Chornobyl Exclusion Zone (sites A18, B18, C18, D18, E18, and F18) and a neighboring non-contaminated area near Slavutych (sites G18, and H18) [7, 8]. The individual total dose rate (ITDR, in µGy/h) absorbed by each tree frog was estimated by assessing internal and external exposure levels, as previously detailed [5, 7]. The ITDR reflects the energy deposited into the frog's body per unit of time. Given that radiation sources are external (the soil, for example) and internal (by ingestion, respiration, ...), the intensity of this deposit is a function of radiation energy, as well as of the organisms' shape, composition, and lifestyle [4, 5, 7]. Missing values for ITDR of 3 individuals were imputed using the site median.

2.1 Transcriptomic Data

The RNA-Seq expression data is published in [3]. For each frog, total RNA was extracted from tibia muscle, and RNA-Seq analysis was performed as previously explained [8]. *De novo* assembly of transcriptome was performed with Trinity using 3 additional tree frogs captured in non-contaminated sites [8, 14]. Estimations of transcript abundance were mapped against the transcriptome with Bowtie2, and their quantification was done with RSEM [19, 23]. R packages `tximport` and `DESeq2` were used to import the quantification data and assemble the RNA-Seq pseudo count matrix [25, 33]. Filtering was applied to remove genes with an expression quantified to 0 for all samples. Given the characteristic mean-variance dependence in RNA-Seq count data, a variance-stabilizing transformation (VST) was applied to the count matrix, producing an approximately homoskedastic matrix with normalized and log-transformed counts [1].

2.2 Genetic Distance

Interindividual genetic distances were obtained by computing the Euclidean distance between SNP genotypes with R package `vcfR` [17]. Genetic distances were summarized into a 3-level categorical variable by performing hierarchical clustering with complete linkage method.

3 Methods

Here, we present the models we used for the correction of batch effects. In the context of our tree frog study, we then describe the exploratory analyses that led to identifying the sources of unwanted variation, the application of BECAs, and the qualitative and quantitative assessment of their results. Computations were performed using R.

3.1 Batch Effect-Correction Algorithms

In the presence of confounding factors, BECAs can be needed to remove the unwanted sources of variation before performing further analyses. Here, we describe three popular batch effect-removal strategies. Linear regression on the batch variable followed by extraction of residuals and ComBat-seq assume that batch factors are known and correct their effect. SVA is used to estimate the unwanted sources of expression heterogeneity.

Linear Residualization Probably the most straightforward strategy when dealing with batch effects is to regress out (or residualize) the batch effects in a pre-processing step. This is done by adjusting a simple (or multiple) linear regression on each feature, using one or more batch variables as regressors, and extracting the residuals [11]. This approach allows the inclusion of several confounders of varied types in the correction process, assuming that batch factors have additive effects. Let $z^{(1)}, \dots, z^{(b)}$ denote b factors, either categorical batch effects or other continuous covariates.

In the classic scenario where there is only one categorical batch factor, removing the batch effects-associated variability using a linear model amounts to performing batch mean-centering (BMC) [32]. BMC adjustment leads to the average value of each variable being zero within each batch.

When batch effects and variable of interest have an unbalanced distribution, removing batch effects can lead to the suppression of actual between-group variation [29]. To avoid over-correction and loss of meaningful biological signal, it is possible to model the effects of the outcome y in the correction step to preserve its influence on the data, as is done by other tools [16, 30, 36, 40]. Under the assumption that outcome-associated effects are additive to batch effects, the expression of gene g for sample $i \in \{1, \dots, n\}$ is modelled as such:

$$\begin{aligned} X_{ig} &= \hat{\beta}_{g0} + \hat{\beta}_{g1} z_i^{(1)} + \cdots + \hat{\beta}_{gb} z_i^{(b)} + \hat{\beta}_{g(b+1)} y_i + \varepsilon_{ig} \\ &= \hat{\beta}_g^\top z_i + \varepsilon_{ig} \end{aligned} \tag{1}$$

with $z_i = (1, z_i^{(1)}, \dots, z_i^{(b)}, y_i)^\top$ containing the batch variables and the variable of interest y, used as regressors, $\hat{\beta}_g = (\hat{\beta}_{g0}, \hat{\beta}_{g1}, \dots, \hat{\beta}_{gb}, \hat{\beta}_{g(b+1)})^\top$ the vector of coefficients estimated by Ordinary Least Squares, and ε_{ig} the residual part. The expression profile is then corrected for the batch effects:

$$\begin{aligned} X_{ig}^\star &= X_{ig} - (\hat{\beta}_{g0} + \hat{\beta}_{g1} z_i^{(1)} + \cdots + \hat{\beta}_{gb} z_i^{(b)}) \\ &= \varepsilon_{ig} + \hat{\beta}_{g(b+1)} y_i \end{aligned} \tag{2}$$

This is also the spirit of the function `removeBatchEffects` from R package `limma` which allows removing the effects of batches and other continuous covariates, while preserving the effect of a condition variable on the log-transformed counts [30].

ComBat-seq ComBat-seq is designed for RNA-Seq count data and is a descendant of ComBat, a widely popular tool for correcting batch effects from known sources [16, 40]. ComBat-seq uses a gene-wise negative binomial regression model to estimate batch effects on the mean and dispersion parameters. After regression parameters are estimated using the raw count matrix, a new batch-corrected distribution is computed and used to adjust the count matrix by mapping quantiles. The biological condition of interest can be included in the modeling to protect between-group differences. We used ComBat-seq both with and without providing ITDR in the covariate model matrix. ComBat-seq is part of the R package `sva` [20].

Surrogate Variable Analysis Unknown sources of variability present in the datasets can be uncovered using SVA [22]. Unobserved variation sources are identified and estimated as follows: after the expression dataset is residualized on the primary variable of interest, the remaining sources of expression heterogeneity are extracted by eigendecomposition and translated into surrogate variables (SVs). These estimated batches can later be used as covariates in subsequent analyses, such as differential expression modeling, or can be adjusted for by residualization prior to visualization algorithms such as Principal Component Analysis (PCA). SVA is implemented in the namesake R package [20]. Function `num.sv` allows the estimation of the number of SVs to be calculated with `sva`.

3.2 Application to the Tree Frog RNA-Seq Study

In the context of the transcriptomic data analysis of Chornobyl tree frogs, we considered the batch effects stemming from the site of collection and the different genetic backgrounds. We also computed and studied surrogate batch variables.

Batch Effect Assessment We performed PCA on the gene expression dataset to extract factors that reflect the largest sources of variability in the data. Data exploration of the first components allowed visualizing the effect of the confounders. Indeed, we expect that principal components will correlate strongly with factors that have the most impact on gene expression. PCA was computed on the VST-transformed count matrix using R function `prcomp`.

Following published batch-effect assessment methodology [21], we tested the association between the 5 first principal components and the qualitative covariates by analysis of variance (ANOVA). We show the rank of the principal component presenting the strongest correlation with the batch using the R-squared value, denoted R^2. R^2 is interpreted as the percentage of variation in the principal component explained by the batch effects.

Similarly, the strength of confounding between categorical batch variables and total dose rate was assessed by generalized R^2 [21]. A correlation of 0 indicates orthogonality between batch and radiocontamination effects, i.e., a low confounding level, while a value close to 1 indicates severe to complete confounding.

Batch Effect Correction

Targeted variables The collection site was primarily considered as the batch variable to target. Further investigation led to the creation of groups based on genetic distance, which were also used as a batch factor. Lastly, a Surrogate Variable Analysis was carried out to estimate 3 SVs, as advised by `num.sv`.

Correction techniques For all candidate variables, we applied batch effect-adjustment algorithms, which returned batch-adjusted matrices. Linear residualization was performed on the VST-transformed count matrix. We combined this method with our three approaches: correction aimed at the site, the genetic group, or the SVs. ComBat-seq was used on the pseudo count matrix, using either the site or the genetic group, and the adjusted matrices were then transformed via VST. ComBat-seq could not be used to remove the effects associated with the SVs, as it can only take one batch variable.

For each experiment, we tested the impact of including or not of the variable of interest in the correction step. In the residualization method, this is done by adding the dose rate as a regressor. In ComBat-seq, we included the dose rate in the covariate model matrix. Table 1 recapitulates the batch effect-adjustment approaches used.

Table 1: Batch effect-correction strategies

Strategy	Targeted batch	BECA
1	Site	Linear residualization
2	Site	ComBat-seq
3	Genetic group	Linear residualization
4	Genetic group	ComBat-seq
5	Surrogate variables	SVA & Linear residualization

Performance Evaluation

Visualization and Summary Statistics We performed PCA on the batch effect-adjusted matrices for performance evaluation purposes. We used visual comparisons by plotting individuals in the first factorial plane. The association of

principal components with the radiation level was tested using the absolute value of Pearson's correlation to check the conservation of relevant biological information.

Functional Enrichment Analysis We performed a biologically informed comparison of batch-adjustment strategies by studying the functional classes of genes that were differentially regulated among individuals [26]. Using sparse principal component analysis (sPCA), we were able to select features (genes) that capture a maximum amount of variance across the radiocontamination gradient for all batch-adjusted datasets. Indeed, with the addition of an L1-penalty, sPCA allows the parametrization of the number of non-zero weights associated with the variables to form principal components [37]. Using function `SPC` from R package `PMA`, we applied sPCA on the variance-stabilized counts and the batch-adjusted matrices. For each dataset, the amount of sparsity in the weight vector was adjusted to yield around 400 non-zero weighted genes in the first component.

A list of stable genes was identified by replicating sPCA on 1000 bootstrap samples. The most stable genes correspond to those selected in more than 70% of cases in either the first or second component or in both. Bootstrap-iterating the computing of sPCA reveals the robustness of the feature selection in regard to sampling. We chose to study the first two components to capture the most important directions of variation in the expression. Stable gene names were then mapped to UniProt IDs for a GO term enrichment analysis performed with R packages `clusterProfiler` [38] and `enrichplot` [38] for visualization. This enrichment analysis allowed us to study the biological processes over-represented in the list of deregulated genes.

4 Results

We examine the confounding effect of the collection site and the impact of its correction using ComBat-seq and residualization. We then present our second approach based on the correction of genetic stratification of the populations. We also study the results of SVA-based correction, which we use to identify potentially undiscovered sources of unwanted variation. Finally, batch effect-mitigation strategies targeting the genetic group and SVs are further compared through functional enrichment analysis.

4.1 Confounding Effect of the Collection Site

The tree frogs were collected in 8 geographical sites to gather samples across a continuum of radiation-pollution levels. Their distribution, both inside the CEZ and outside, is shown in Figure 1a. Tree frogs living in different sites exhibit distinct expression patterns, as is revealed by PCA in Figure 1c. Because of the distribution of radiation pollution in the Chornobyl area, individuals captured in one site have likely been exposed with a similar intensity, see Figure 1b. The high confounding between the site and the variable of interest reflects this.

However, site-associated variation in expression poorly reflects the radiocontamination variation, raising the need to distinguish between site and radiation exposure's effects.

We applied batch-correction techniques targeting the site, and we obtained batch-corrected matrices. In Figure 1d, we correlate principal components obtained on the unadjusted and site-adjusted matrices. We handled the site effect using linear residualization on the site and ComBat-seq, both with or without including radiation levels in the correction step. Adjusting the dataset without modeling the radiation-associated effect appears to have failed to preserve the biological information of interest since no principal components are correlated with ITDR after correction. On the other hand, batch-mitigation models that included radiation level seem to have brought forward the radiation-correlated variance. However, several studies have raised the issue of the unreliability of BECAs when adjusting for a highly confounded batch while preserving class differences [29]. The concern is that artificial variation in the direction of dose rate variation could be introduced in the "corrected" dataset, resulting in exaggerated significance in downstream analyses. Without ground-truth measurements to test for this issue, we searched for a way to deconvolute the batch effects from the variable of interest.

4.2 Genetic Diversity Treated as Batch Effects

In a second approach, we wondered which parameters, other than radionuclide pollution, could drive the expression heterogeneity among populations. Firstly, all the frogs were captured in wetlands and presumably had the same living conditions regarding food availability, temperature, or habitat type. Secondly, tree frog populations were sampled site by site, visiting one or two sites per day over a 9-day period. Therefore, even with a thought-out protocol, we cannot exclude that the effect of the site may have a technical cause. Lastly, the geographical distance between sites (80km between the two most distant) and the limited dispersal radius of this species could translate into genetic background dissimilarities. Genetic distance between individuals was used to integrate this aspect into the analysis and was translated into a categorical batch variable by hierarchical clustering (Figure 2a). The obtained batch factor is less confounded with ITDR than the site while still being strongly correlated with the third principal component (Table 2, Figure 2b). We note that the principal component that carries the most variance in the direction of radiation contamination is of a lesser rank.

We performed genetic group adjustment with the techniques mentioned previously, and we present the results of PCA applied to the adjusted matrices in Figure 2. Figure 2c illustrates that the first two PCs now do a fine job gathering the expression heterogeneity along the contamination gradient. Individuals still appear to cluster by site, but this is explicable by the dose rate disparity between sites. Regarding the correlation of the PCs with the variable of interest (Figure 2d), it appears that correcting the effect of genetic structure protects the radiation-associated variation even without integrating the variable in the

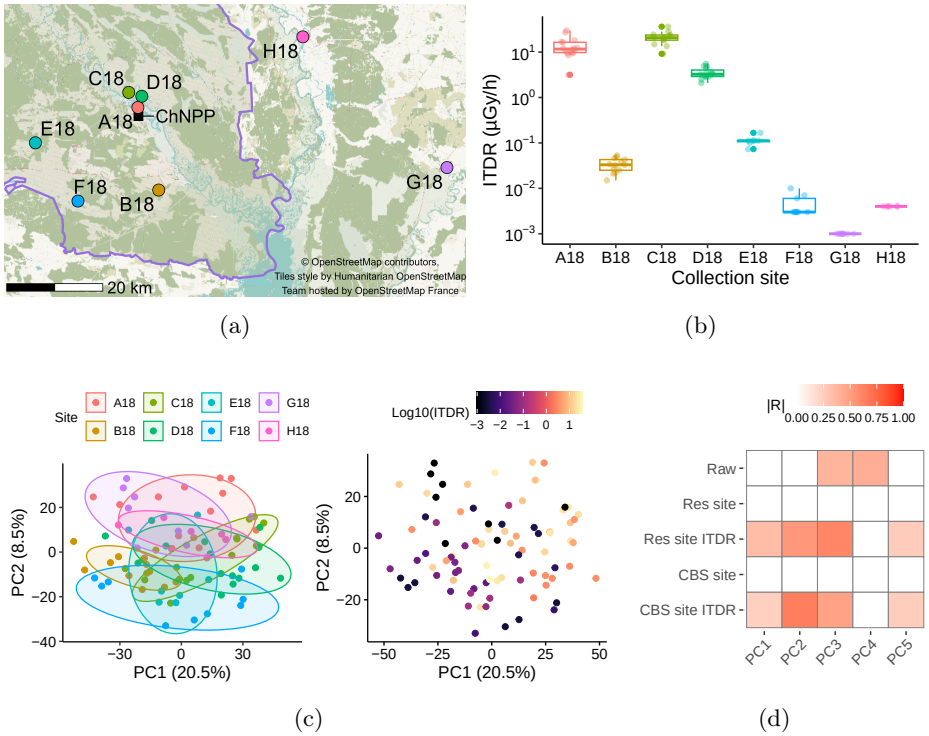


Fig. 1: Confounding effect of the collection site: visualization and correction **a.** Geographical location of tree frog-collection sites inside the CEZ (purple line) and in the Slavutych region (ChNPP = Chornobyl nuclear power plant) **b.** Dose rate distribution among sites **c.** PCA of raw RNA-seq count matrix with individuals colored by collection site or dose rate **d.** Absolute value of the correlation between radiation level and PCs of site-corrected count matrices (CBS = ComBat-seq, Res = Residualization, ITDR = ITDR was included in the batch-correction step to preserve its effects)

correction step. Here, we identified a previously unconsidered batch variable by interrogating another data modality. In the next step, we studied the possibility of other unknown sources of unwanted variability.

4.3 Search of Other Possible Confounders with Surrogate Variable Analysis

We applied SVA to diagnose the sources of variation in the VST-transformed expression dataset. After capturing and extracting the linear relationship between expression and radiation exposure, 3 surrogate variables (SVs) were identified as significant sources of systemic variation. In Figure 3a, we compare the SVs with the covariates we know of and which have been measured, by computing the

Table 2: Confounding effects of site and genetic distance

	Batch variables		Variable of interest
	Site of collection	Genetic distance group	Radiocontamination level (ITDR)
Confounding with ITDR	0.80	0.08	×
Rank of PC with highest correlation (R^2)	3 (0.72)	3 (0.35)	4 (0.42)

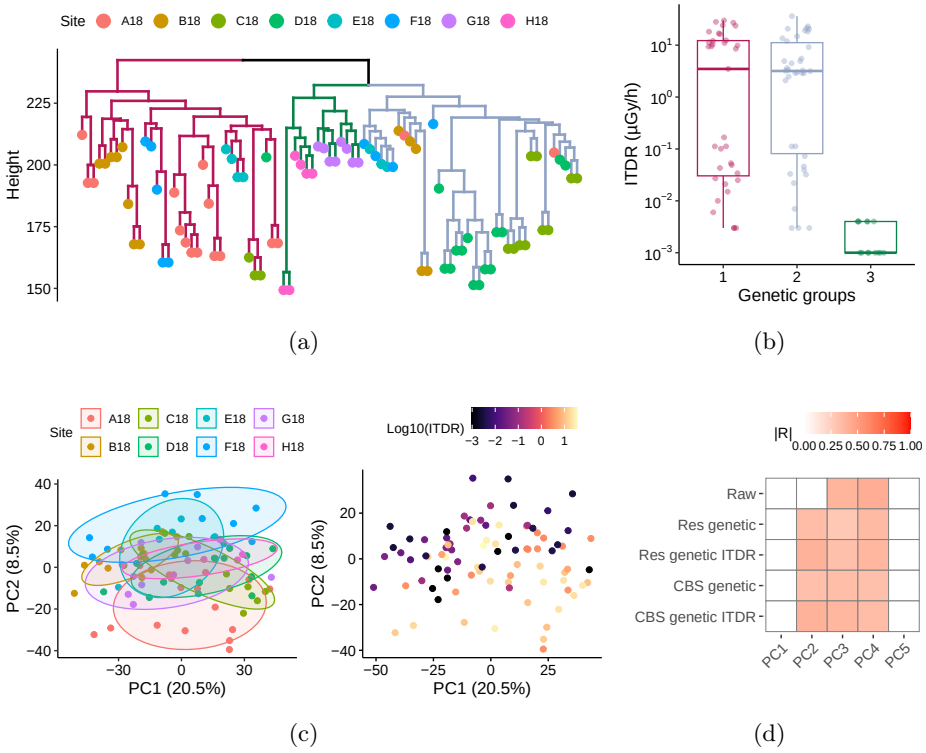
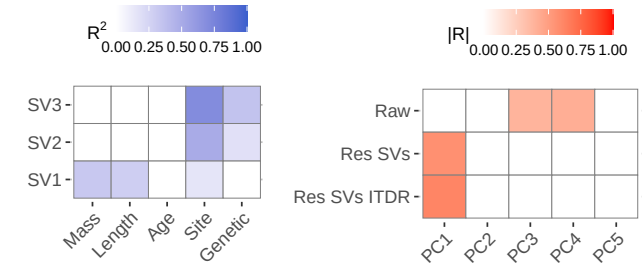


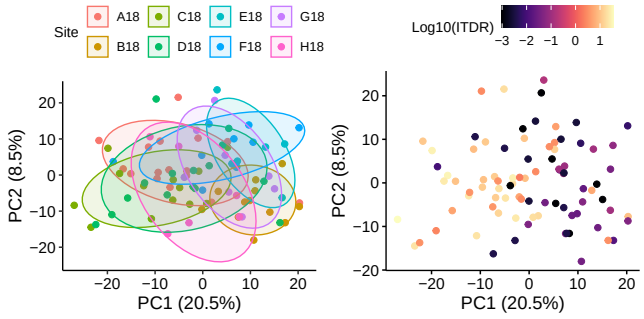
Fig. 2: Construction of genetic population grouping and impact of its correction **a.** Hierarchical clustering on genetic distance **b.** Dose rate distribution among genetic groups **c.** PCA after correction of genetic group with ComBat-seq **d.** Absolute value of the correlation between radiation level and PCs of genetic group-corrected count matrices (CBS = ComBat-seq, Res = Residualization, ITDR = ITDR was included in the batch-correction step to preserve its effects)

adjusted R^2 of the linear regression of SVs on covariates. The constructed SVs correlate with known potential covariates, such as the mass and length (SV1) and the site and genetic group for SVs 2 and 3. None of the estimated SVs capture age-related expression patterns.

As we expected from the primary exploration of the dataset, the collection site appears to have the strongest influence among known covariates. SVs capture expression trends that correlate with the site factor, even after radiation effects were residualized by SVA, which confirms that radionuclide pollution is not the only driver of differences between sites. We applied linear residualization to adjust for the estimated SVs, both with and without modeling ITDR effects, and present results from the PCA on the SVA-corrected matrix in Figures 3b- 3c. The first principal component significantly correlates with radiocontamination levels and allows us to distinguish well between the highest and lowest dose rate groups.



(a) (b)



(c)

Fig. 3: Surrogate variables: connection to known covariates and correction **a.** Correlation between estimated SVs and covariates **b.** Absolute value of the correlation between radiation level and PCs of SVs-corrected count matrices (Res = Residualization, ITDR = ITDR was included in the batch-correction step to preserve its effects) **c.** PCA after correction of SVs-corrected by residualization

4.4 Biological Interpretation

Correction methods targeting either the genetic structure or the surrogate variable both appear to have removed confounding effects. We now want to compare the two approaches through functional enrichment analysis. Using the raw and batch-adjusted datasets, we performed sPCA coupled to bootstrap to identify stably selected genes that carry the most variance in each dataset. Enrichment analysis then searches for the biological pathways in which the selected genes participate, thus indicating the deregulation of expression in the associated biological functions.

Table 3 presents the number of stably selected genes for each dataset and the number of Gene Ontology (GO) terms significantly enriched. We note that highly similar numbers of processes were recovered in the raw and ComBat-seq corrected datasets. This good performance of the uncorrected matrix could be explained by the robustness of sparse PCA compared to standard PCA, allowing the true biological signal to be distinguished from noise even without batch correction (see Appendix A). ComBat-seq-adjusted matrices yield approximately as many stable genes as the raw matrix, while residualized matrices return considerably less. Residualization coupled with SVA leads to the selection of very few genes, which suggests over-correction and removal of relevant information. Also, for the SVA method, we note that including the radiation levels in the correction step doubles the number of GO terms identified, whereas it does not impact the count for the other methods. This could signify residual confounding between the SVs and the variable of interest.

Table 3: Feature selection and Gene Ontology terms enrichment before and after batch correction

Correction method	Number of stably selected genes	Number of enriched GO terms
Raw	642	77
Residualization on genetic group	363	56
Residualization on genetic group with preservation of ITDR effects	441	58
ComBat-seq on genetic group	660	80
ComBat-seq on genetic group with preservation of ITDR effects	653	77
Residualization on SVs	33	24
Residualization on SVs with preservation of ITDR effects	95	48

In Figure 4, we present the most significant deregulated pathways identified after the different correction methods. The nodes, which represent biological

processes, are linked according to the sharing of genes involved in the processes. Pathways recovered by the various techniques do not perfectly overlap. Notably, GO terms connected to cellular respiration or mitochondrial activity are enriched in datasets corrected for genetic structure and the raw dataset but disappear after SVA correction. Conversely, deregulated pathways identified consecutively to SVA correction are heavily biased toward muscle processes. We find that correction methods applied to the genetic groups preserved more relevant biological information than the residualization of SVs.

Overall, the study of deregulated biological processes highlights effects on energy metabolism (GO terms such as "ATP metabolic process" or "energy derivation by oxidation of organic compounds") and muscle processes (GO terms "muscle system process" or "muscle tissue development"). Perturbation in these pathways is consistent with impacts of exposure to low-dose radiation reported in studies of other organisms [18, 27, 28].

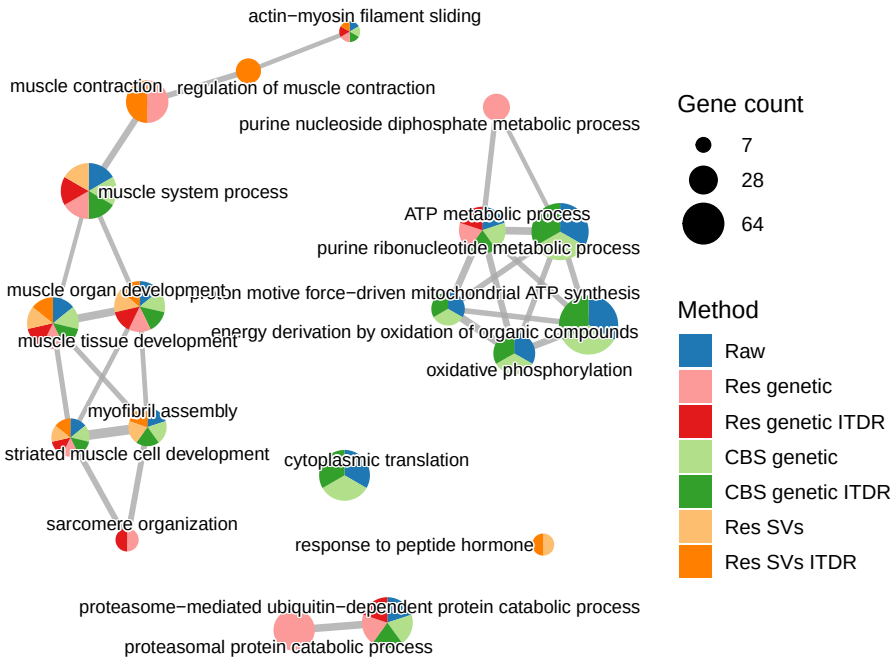


Fig. 4: Enriched Gene Ontology-terms network for each correction strategy considered (CBS = ComBat-seq, Res = Residualization, ITDR = ITDR was included in the batch-correction step to preserve its effects)

5 Discussion

Classical BECAs used to correct for the confounded effect of the site either resulted in the suppression of valuable biological variation or in the lack of reliability of the adjusted dataset when including the radiation level in the modeling. Instead, we successfully deconvoluted the batch from the variable of interest by including the genetic distance between individuals as a new variable to correct. The computed genetic batch reflects population structure and observed admixture between populations [8]. In the context of the CEZ, genetic distance between individuals could be a testimony of inherited mutations due to historic radiation pollution. Here, we hypothesize that removing population structure effects, which we associate with historical exposure, allowed us to better investigate the impact of current exposure to low-dose radiation.

When its origin is unknown, the confounder is represented by a proxy variable that encompasses the true source of variation. For example, in the case of technical batch effects, the day and group of processing can be the only accessible information [21]. Regarding the site effects, we assume that their origin is not technical, but we lack the information to assert it with certainty. Nonetheless, the consistency between the genetic structure of CEZ populations and their geographical distribution makes the genetic batch a fit candidate. Also, it is not uncommon to correct for population structure, for instance, in genome-wide association studies [34]. We concede that this approach is modality-dependent and could be less relevant when studying proteomics or metabolomics.

Our experiments illustrated the limitations of ComBat-seq and residualization in a batch-class unbalance design, as was acknowledged by several studies that suggest using BECAs with caution in confounded scenarios [12, 24, 29, 41]. The genetic distance-adjusted datasets appear to be less affected by the integration or not of the variable of interest in the correction step. Feature selection on adjusted datasets followed by functional enrichment showed that relevant biological information was preserved better with ComBat-seq than with residualization. SVA recovered sources of unwanted variation, which is an appealing strategy to detangle the confounding variable from the level of radiation exposure. Residualization of estimated SVs provided satisfying results in the PCA visualization, but the enrichment analysis revealed a substantial loss of information relative to radiation exposure. Indeed, SVA authors warn against correcting SVs in a pre-processing step and instead advise that SVs be used as covariates in following analyses such as differential expression [15].

When batch effects cannot be avoided, and happen to covary with a variable of biological interest, the recommended mitigation strategy is to take into account the batch factor as a covariate directly in the main analysis [29]. For gene expression analysis, this is possible using popular differential expression tools [9, 25, 30]. For all types of data, if the confounder is a categorical variable, multi-group approaches allow integrating batches in a joint factorial analysis [10, 31, 35, 36]. However, in the multi-omics framework, researchers are advised to assess and handle confounding effects in the different modalities before performing the joint analysis [39]. With the development of multi-omics studies, we hope that

more tools will be published that allow accounting for batch effects or other group structures, with already a few examples in the single cell context [2, 6].

6 Conclusion

In the presence of a confounded batch effect, we were able to deconvolute the batch variable from the variable of interest through the integration of additional information. Our correction strategy based on genetic distance allowed us to handle the confounding effect of the frogs' collection site. We were able to identify distinctive changes in gene expression associated with chronic radiation exposure in Chornobyl tree frogs, by implementing this approach (see [8]). Future work will include the development of a new method for the incorporation of confounding factors into multimodal analyses based on the RGCCA framework.

Acknowledgments. EG and CC are supported by PhD grants funded by the French Institute for Radiation Protection and Nuclear Safety (IRSN). S. Gashchack, Y. Gulyaichenko, G. Orizaola, and P. Burraco helped in the field, and S. Gashchack also with measurements of radioactive contamination in tree frogs.

A Appendix: Close-up on the Impact of Sparsity in Gene Selection

To assess how forcing sparsity in the PCA influenced the selected gene list, we also ran a similar approach using standard PCA. We performed PCA on 1000 bootstrap samples of the non-corrected (raw) variance-stabilized matrix. In each model, genes were ranked by the absolute value of their weight, and the top 400 genes were selected from components 1 and 2. Genes stably selected across bootstrap iterations in components 1, 2, or both were submitted to gene functional annotation, as mentioned previously.

Table 4 shows that the genes selected using sparse PCA were more stable across bootstrap samples than with standard PCA. This led to identifying a larger number of deregulated pathways in the uncorrected dataset with sparse PCA than with PCA. In Figure 5, we notice that the alteration of biological processes related to energy metabolism (GO terms "oxidative phosphorylation" or "energy derivation by oxidation of organic compounds") was recovered with sPCA and not with PCA. The identification of pathways typically linked with low-dose radiation, despite the presence of batch effects, suggests that the imposition of weight sparsity in the PCA mitigated the influence of noise.

Table 4: Feature selection approaches and Gene Ontology terms enrichment

Feature selection method	Number of stably selected genes	Number of enriched GO terms
Standard PCA (Raw)	384	52
Sparse PCA (Raw)	642	77

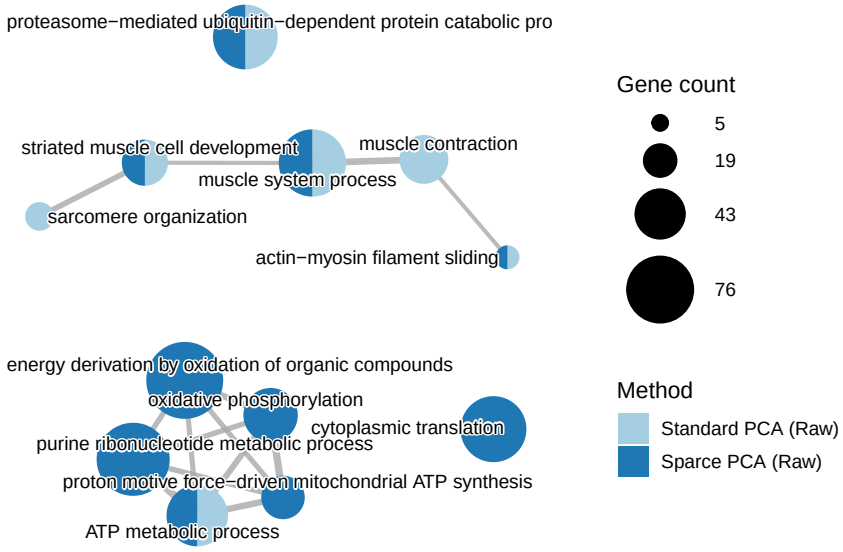


Fig. 5: Enriched Gene Ontology-terms network after feature selection on the uncorrected count matrices

References

1. Anders, S., Huber, W.: Differential expression analysis for sequence count data. Genome Biology **11**(10), R106 (2010). https://doi.org/10.1186/gb-2010-11-10-r106
2. Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J.C., Stegle, O.: MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biology **21**(1), 111 (2020). https://doi.org/10.1186/s13059-020-02015-1
3. Armant, O., Car, C., Frelon, S., Camoin, L.: Population transcriptogenomics highlights impaired metabolism and small population sizes in tree frogs living in the Chernobyl Exclusion Zone (2023), https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE211060
4. Beaugelin-Seiller, K., Jasserand, F., Garnier-Laplace, J., Gariel, J.C.: Modeling radiological dose in non-human species: principles, com-

puterization, and application. Health Physics **90**(5), 485–493 (2006). https://doi.org/10.1097/01.HP.0000182192.91169.ed
5. Burraco, P., Car, C., Bonzom, J.M., Orizaola, G.: Assessment of exposure to ionizing radiation in Chernobyl tree frogs (Hyla orientalis). Scientific Reports **11**, 20509 (2021). https://doi.org/10.1038/s41598-021-00125-9
6. Cao, Z.J., Gao, G.: Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. Nature Biotechnology **40**(10), 1458–1466 (2022). https://doi.org/10.1038/s41587-022-01284-4
7. Car, C., Gilles, A., Armant, O., Burraco, P., Beaugelin-Seiller, K., Gashchak, S., Camilleri, V., Cavalié, I., Laloi, P., Adam-Guillermin, C., Orizaola, G., Bonzom, J.M.: Unusual evolution of tree frog populations in the Chernobyl exclusion zone. Evolutionary Applications **15**(2), 203–219 (2022). https://doi.org/10.1111/eva.13282
8. Car, C., Gilles, A., Goujon, E., Muller, M.L.D., Camoin, L., Frelon, S., Burraco, P., Granjeaud, S., Baudelet, E., Audebert, S., Orizaola, G., Armengaud, J., Tenenhaus, A., Garali, I., Bonzom, J.M., Armant, O.: Population transcriptogenomics highlights impaired metabolism and small population sizes in tree frogs living in the Chernobyl Exclusion Zone. BMC biology **21**(1), 164 (2023). https://doi.org/10.1186/s12915-023-01659-2
9. Chen, Y., Chen, L., Lun, A.T.L., Baldoni, P.L., Smyth, G.K.: edger 4.0: powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets. bioRxiv (2024). https://doi.org/10.1101/2024.01.21.576131
10. Eslami, A., Qannari, E.M., Kohler, A., Bougeard, S.: Algorithms for multi-group PLS. Journal of Chemometrics **28**(3), 192–201 (2014). https://doi.org/10.1002/cem.2593
11. García, C.B., Salmerón, R., García, C., García, J.: Residualization: justification, properties and application. Journal of Applied Statistics **47**(11), 1990–2010 (2020). https://doi.org/10.1080/02664763.2019.1701638
12. Goh, W.W.B., Wang, W., Wong, L.: Why Batch Effects Matter in Omics Data, and How to Avoid Them. Trends in Biotechnology **35**(6), 498–507 (2017). https://doi.org/10.1016/j.tibtech.2017.02.012
13. Goh, W.W.B., Yong, C.H., Wong, L.: Are batch effects still relevant in the age of big data? Trends in Biotechnology **40**(9), 1029–1040 (2022). https://doi.org/10.1016/j.tibtech.2022.02.005
14. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A.: Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nature biotechnology **29**(7), 644–652 (2011). https://doi.org/10.1038/nbt.1883
15. Jaffe, A.E., Hyde, T., Kleinman, J., Weinbergern, D.R., Chenoweth, J.G., McKay, R.D., Leek, J.T., Colantuoni, C.: Practical impacts of genomic data "cleaning" on biological discovery using surrogate variable analysis. BMC Bioinformatics **16**(1), 372 (2015). https://doi.org/10.1186/s12859-015-0808-5
16. Johnson, W.E., Li, C., Rabinovic, A.: Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics **8**(1), 118–127 (2007)
17. Knaus, B.J., Grünwald, N.J.: vcfr: a package to manipulate and visualize variant call format data in R. Molecular Ecology Resources **17**(1), 44–53 (2017). https://doi.org/10.1111/1755-0998.12549

18. Kostyuk, S.V., Proskurnina, E.V., Konkova, M.S., Abramova, M.S., Kalianov, A.A., Ershova, E.S., Izhevskaya, V.L., Kutsev, S.I., Veiko, N.N.: Effect of Low-Dose Ionizing Radiation on the Expression of Mitochondria-Related Genes in Human Mesenchymal Stem Cells. International Journal of Molecular Sciences **23**(1), 261 (2021). https://doi.org/10.3390/ijms23010261
19. Langmead, B., Salzberg, S.L.: Fast gapped-read alignment with Bowtie 2. Nature methods **9**(4), 357–359 (2012). https://doi.org/10.1038/nmeth.1923
20. Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E., Storey, J.D.: The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics **28**(6), 882–883 (2012). https://doi.org/10.1093/bioinformatics/bts034
21. Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K., Irizarry, R.A.: Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics **11**(10), 733–739 (2010). https://doi.org/10.1038/nrg2825
22. Leek, J.T., Storey, J.D.: Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. PLoS Genetics **3**(9), e161 (2007). https://doi.org/10.1371/journal.pgen.0030161
23. Li, B., Dewey, C.N.: RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics **12**(1), 323 (2011). https://doi.org/10.1186/1471-2105-12-323
24. Li, T., Zhang, Y., Patil, P., Johnson, W.E.: Overcoming the impacts of two-step batch effect correction on gene expression estimation and inference. Biostatistics **24**(3), 635–652 (2023). https://doi.org/10.1093/biostatistics/kxab039
25. Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology **15**(12), 550 (2014). https://doi.org/10.1186/s13059-014-0550-8
26. Martinelli, F., Reagan, R., Uratsu, S., Phu, M., Albrecht, U., Zhao, W., Davis, C., Bowman, K., Dandekar, A.: Gene Regulatory Networks Elucidating Huanglongbing Disease Mechanisms. PloS one **8**, e74256 (2013). https://doi.org/10.1371/journal.pone.0074256
27. Murat El Houdigui, S., Adam-Guillermin, C., Armant, O.: Ionising Radiation Induces Promoter DNA Hypomethylation and Perturbs Transcriptional Activity of Genes Involved in Morphogenesis during Gastrulation in Zebrafish. International Journal of Molecular Sciences **21**(11), 4014 (2020). https://doi.org/10.3390/ijms21114014
28. Murat El Houdigui, S., Adam-Guillermin, C., Loro, G., Arcanjo, C., Frelon, S., Floriani, M., Dubourg, N., Baudelet, E., Audebert, S., Camoin, L., Armant, O.: A systems biology approach reveals neuronal and muscle developmental defects after chronic exposure to ionising radiation in zebrafish. Scientific Reports **9**(1), 20241 (2019). https://doi.org/10.1038/s41598-019-56590-w
29. Nygaard, V., Rødland, E.A., Hovig, E.: Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. Biostatistics **17**(1), 29–39 (2016). https://doi.org/10.1093/biostatistics/kxv027
30. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., Smyth, G.K.: limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research **43**(7), e47 (2015). https://doi.org/10.1093/nar/gkv007

31. Rohart, F., Eslami, A., Matigian, N., Bougeard, S., Lê Cao, K.A.: MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. BMC Bioinformatics **18**(1), 128 (2017). https://doi.org/10.1186/s12859-017-1553-8
32. 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 (2008). https://doi.org/10.1186/1755-8794-1-42
33. Soneson, C., Love, M.I., Robinson, M.D.: Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research **4**, 1521 (2016). https://doi.org/10.12688/f1000research.7563.2
34. Sul, J.H., Martin, L.S., Eskin, E.: Population structure in genetic studies: Confounding factors and mixed models. PLoS Genetics **14**(12), e1007309 (2018). https://doi.org/10.1371/journal.pgen.1007309
35. Tenenhaus, A., Tenenhaus, M.: Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis. European Journal of Operational Research **238**(2), 391–403 (2014). https://doi.org/10.1016/j.ejor.2014.01.008
36. Wang, Y., Lê Cao, K.A.: PLSDA-batch: a multivariate framework to correct for batch effects in microbiome data. Briefings in Bioinformatics **24**(2), bbac622 (2023). https://doi.org/10.1093/bib/bbac622
37. Witten, D.M., Tibshirani, R., Hastie, T.: A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics **10**(3), 515–534 (2009). https://doi.org/10.1093/biostatistics/kxp008
38. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., Yu, G.: clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation **2**(3), 100141 (2021). https://doi.org/10.1016/j.xinn.2021.100141
39. Yu, Y., Zhang, N., Mai, Y., Ren, L., Chen, Q., Cao, Z., Chen, Q., Liu, Y., Hou, W., Yang, J., Hong, H., Xu, J., Tong, W., Dong, L., Shi, L., Fang, X., Zheng, Y.: Correcting batch effects in large-scale multiomics studies using a reference-material-based ratio method. Genome Biology **24**(1), 201 (2023). https://doi.org/10.1186/s13059-023-03047-z
40. Zhang, Y., Parmigiani, G., Johnson, W.E.: ComBat-seq: batch effect adjustment for RNA-seq count data. NAR genomics and bioinformatics **2**(3), lqaa078 (2020). https://doi.org/10.1093/nargab/lqaa078
41. Zhou, L., Chi-Hau Sue, A., Bin Goh, W.W.: Examining the practical limits of batch effect-correction algorithms: When should you care about batch effects? Journal of Genetics and Genomics **46**(9), 433–443 (2019). https://doi.org/10.1016/j.jgg.2019.08.002