



HAL
open science

Handling confounding factors in analyzing the transcriptomic data from Chornobyl tree frogs

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

► **To cite this version:**

Elen Goujon, Olivier Armant, Jean-Marc Bonzom, Arthur Tenenhaus, Imène Garali. Handling confounding factors in analyzing the transcriptomic data from Chornobyl tree frogs. 24th Journées Ouvertes en Biologie, Informatique et Mathématiques, Jun 2023, Plouzané, France. . irsn-04730233

HAL Id: irsn-04730233


<https://irsn.hal.science/irsn-04730233v1>

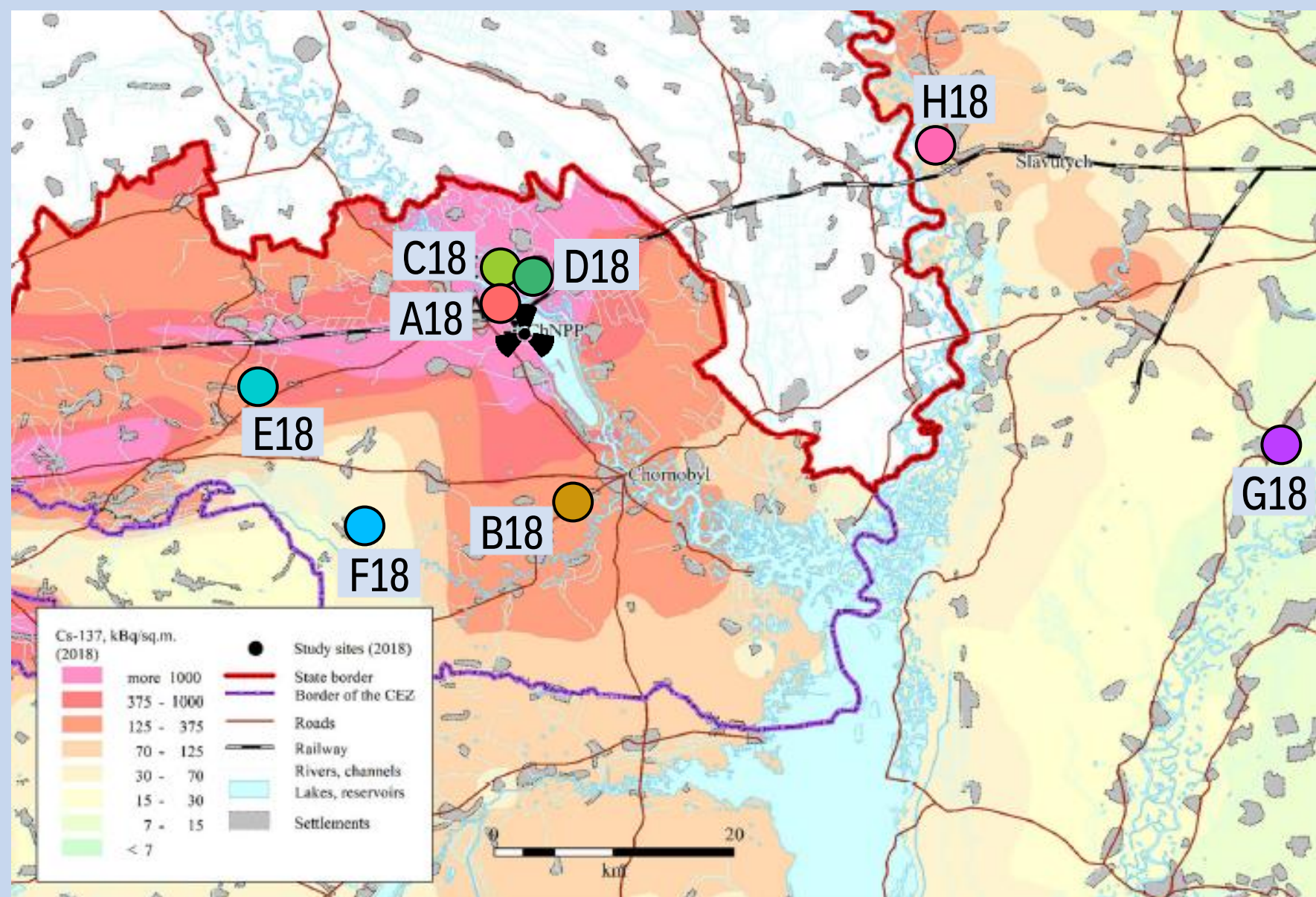
Submitted on 10 Oct 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.


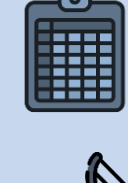



I. Context


 Nuclear power plant accident in Chernobyl, Ukraine (1986): release of ionizing radiation and environmental contamination

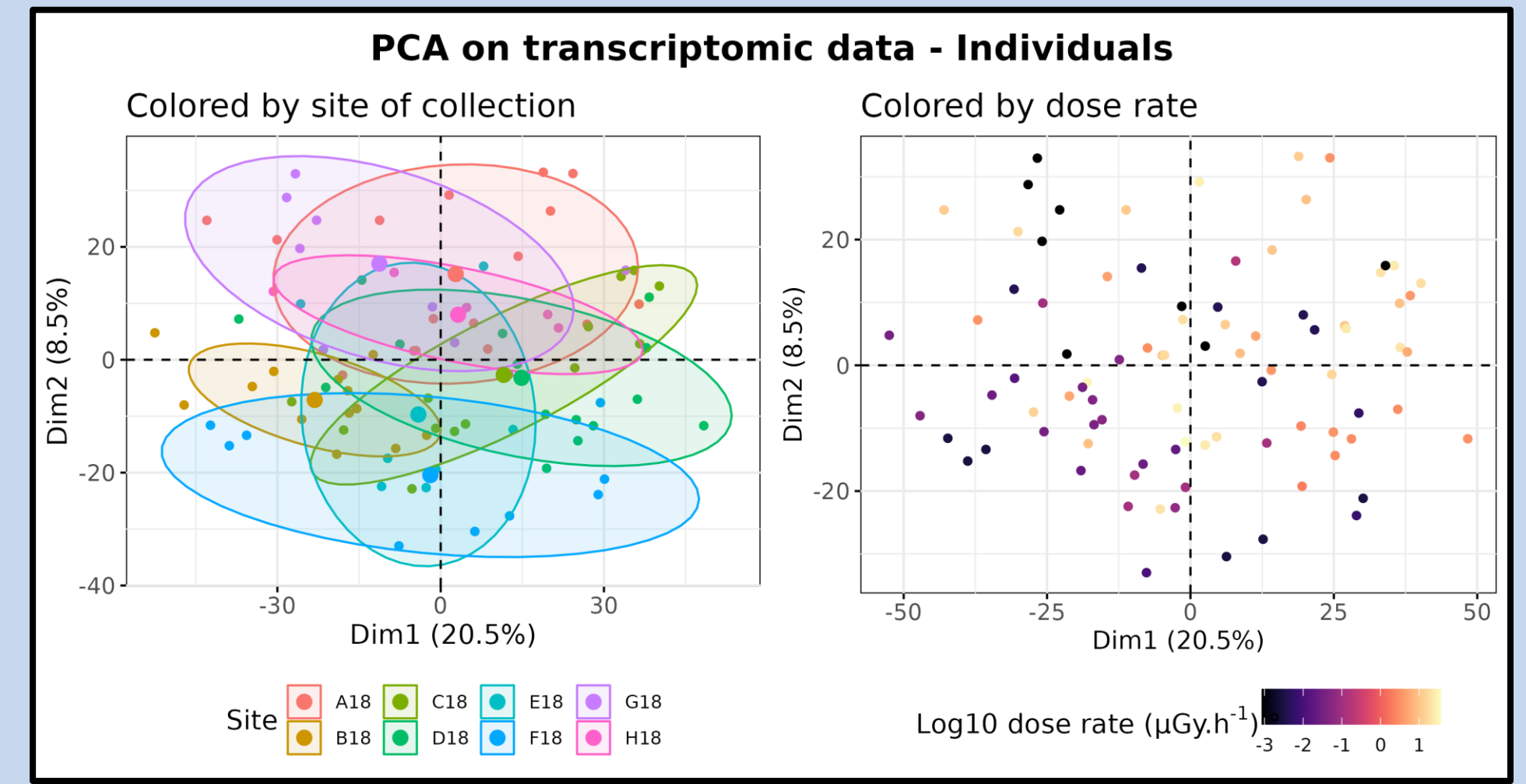


Sampling locations of Eastern tree frog *Hyla orientalis* populations in the exclusion zone in 2018[1,2]

 Collection of tree frog population samples

- 87 males from 8 sites, including 6 within the Chernobyl Exclusion Zone (CEZ)
- Exposed to various levels of contamination
- Omics and non-omics data extraction:
 -  Dosimetry: total dose rate with internal and external contributions
 -  Age, phenotype (mass, dimensions)
 -  Genomics (DNA-seq)
 -  Transcriptomics (RNA-seq)
 -  Proteomics

 Confounding factor for the transcriptomic data analysis



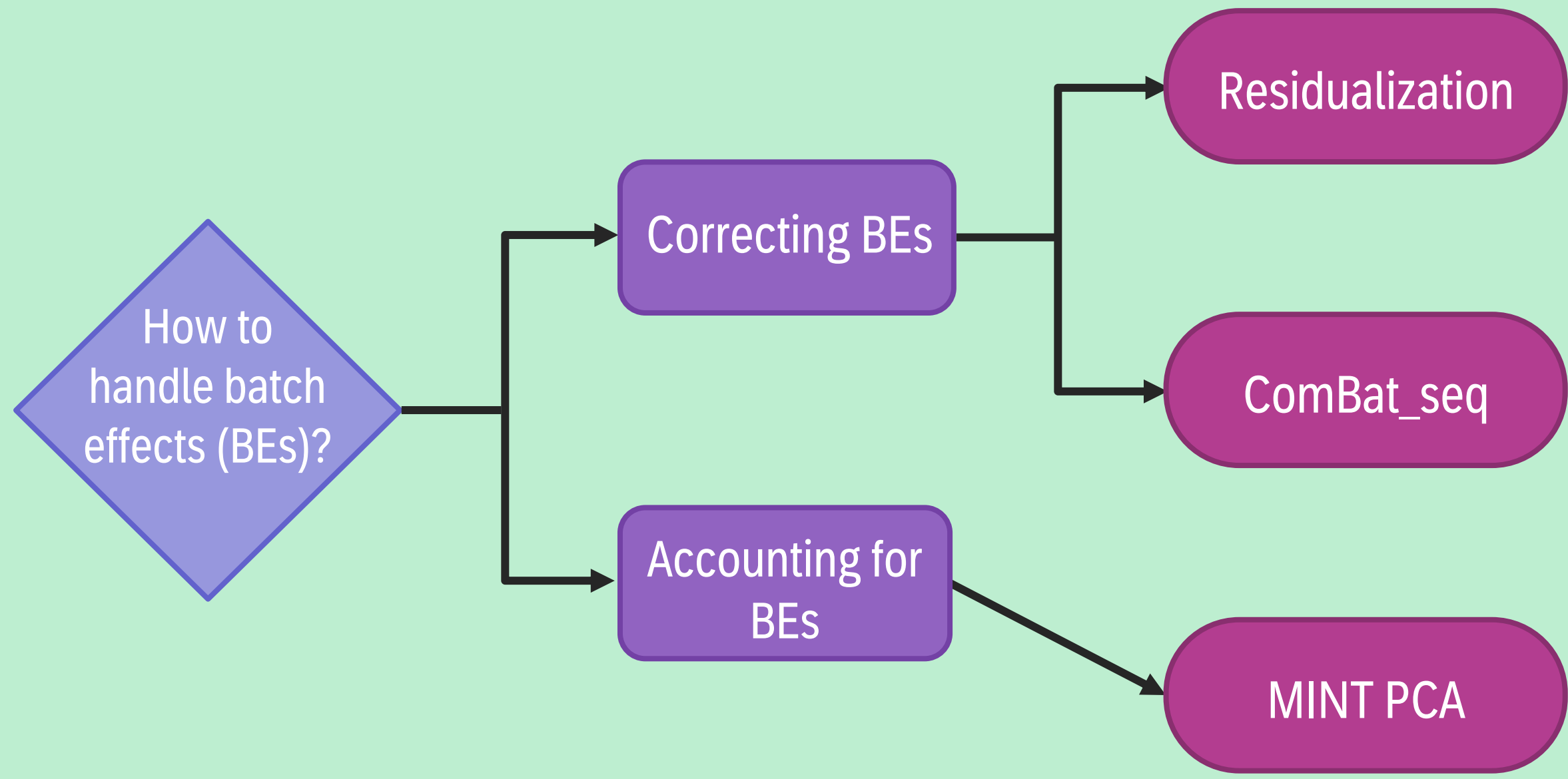
A principal component analysis (PCA) performed on the RNA-seq data reveals the association with the site. The site factor appears to have a stronger role than the dose in explaining the variability in the transcriptomic data.

Objective: Identify specific molecular signatures of exposure to better understand the effects of low-dose radiation

Methodological challenge

→ Handling factors confounding the analysis of the effect of the dose

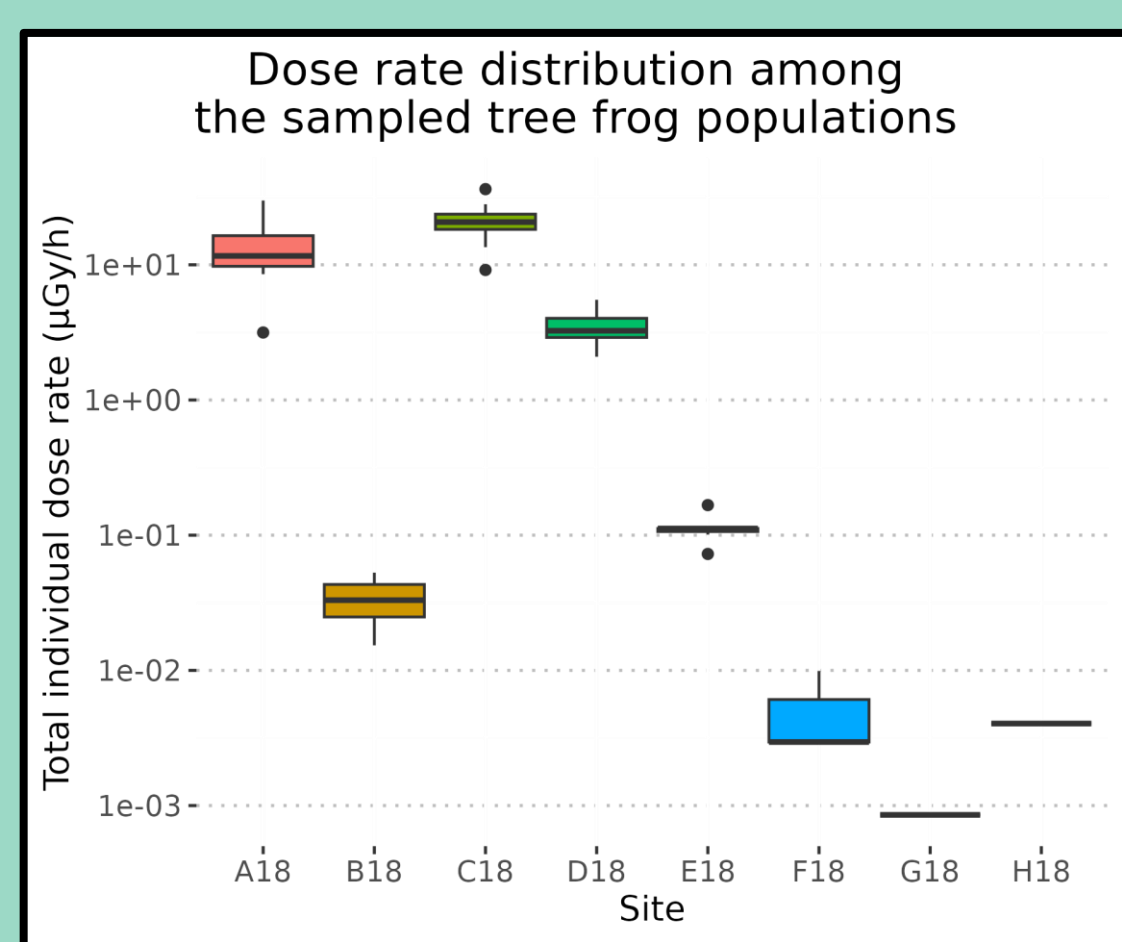
II. Strategies for batch effect mitigation



When to use	How to use in R	Principle	Algorithm
Pre-processing step	Native stats package	Removes linear relationship between variables and batch	For each variable: 1. Fit a linear regression model on the variable using the batch factor as a regressor 2. Extract the residual part
Pre-processing step [3]	sva package	Corrects data distribution to fit a batch-free distribution	1. Estimate the parameters of the negative binomial regression model $y_{gij} \sim NB(\mu_{gij}, \phi_{gi})$ 2. Compute the batch-corrected distributions 3. Adjust data to the batch-corrected distributions by quantile mapping
Integrated method [4]	mixOmics package	Within-group centering and scaling followed by PCA	1. Data is centered and scaled within groups 2. PCA algorithm is performed (based on SVD matrix diagonalization)

III. Controlling for the site effect

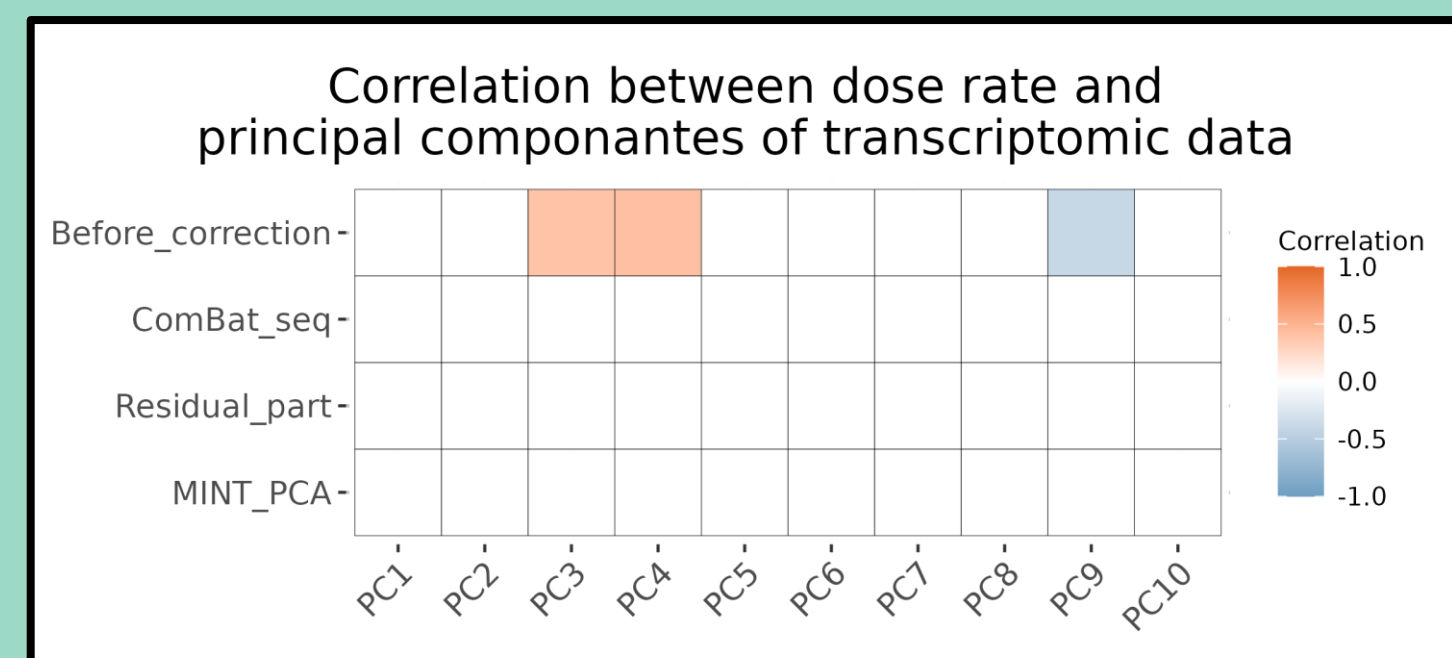
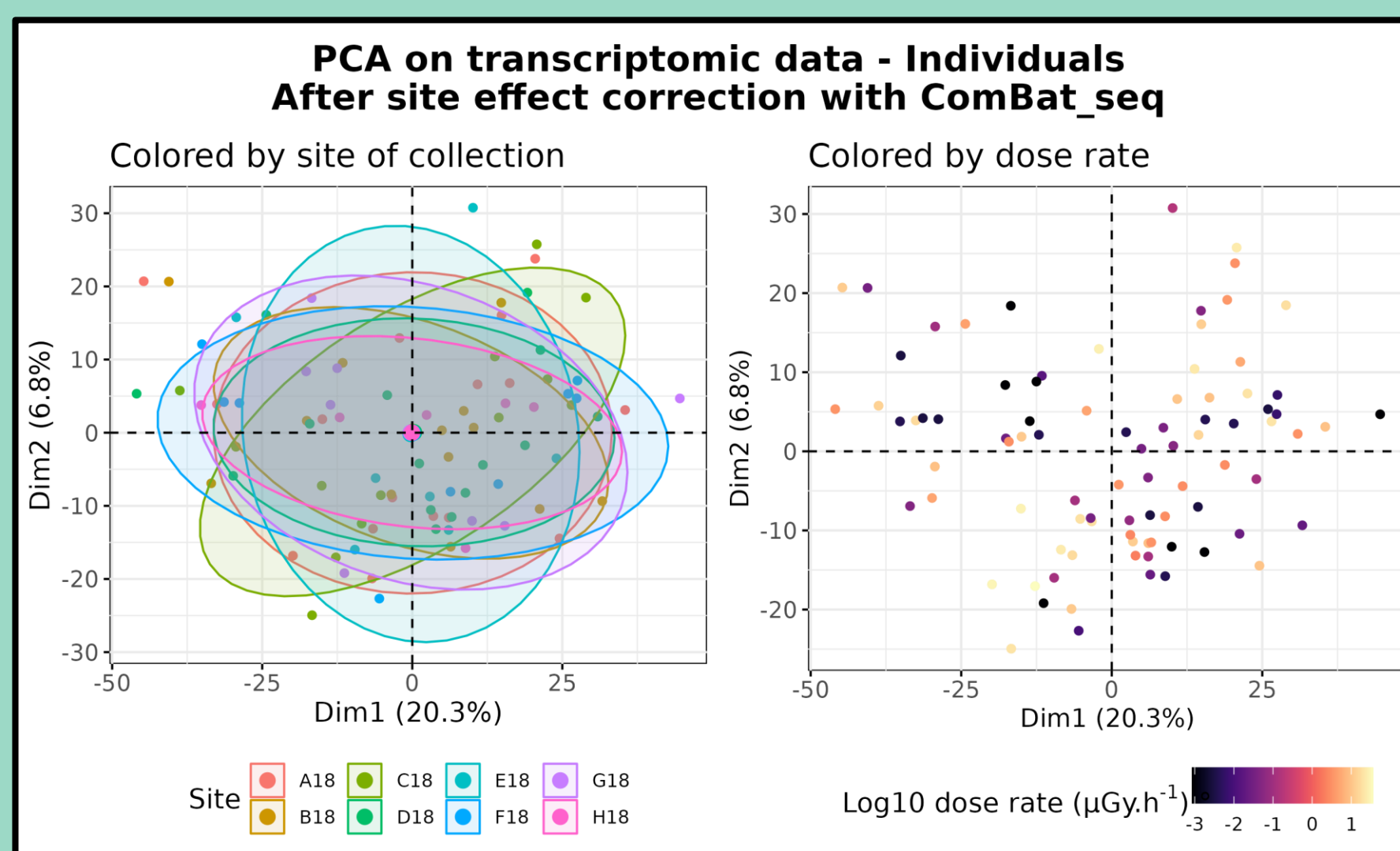
Batch effect mitigation is a common harmonization problem when integrating data obtained across different experiments, individuals or technologies. Methods either account for BEs, by including the batch as a covariate, or correct them by removing the unwanted variation in a pre-treatment step. We applied PCA on the corrected count matrices to compare performances with those of MINT PCA.



Severe confoundedness between site and dose: a challenge

The site effect is particularly complex to remove as it is severely confounded with the dose rate. The correlation between principal components and dose rate was computed to test the preservation of dose-related biological information. This metric was used to compare methods quantitatively.

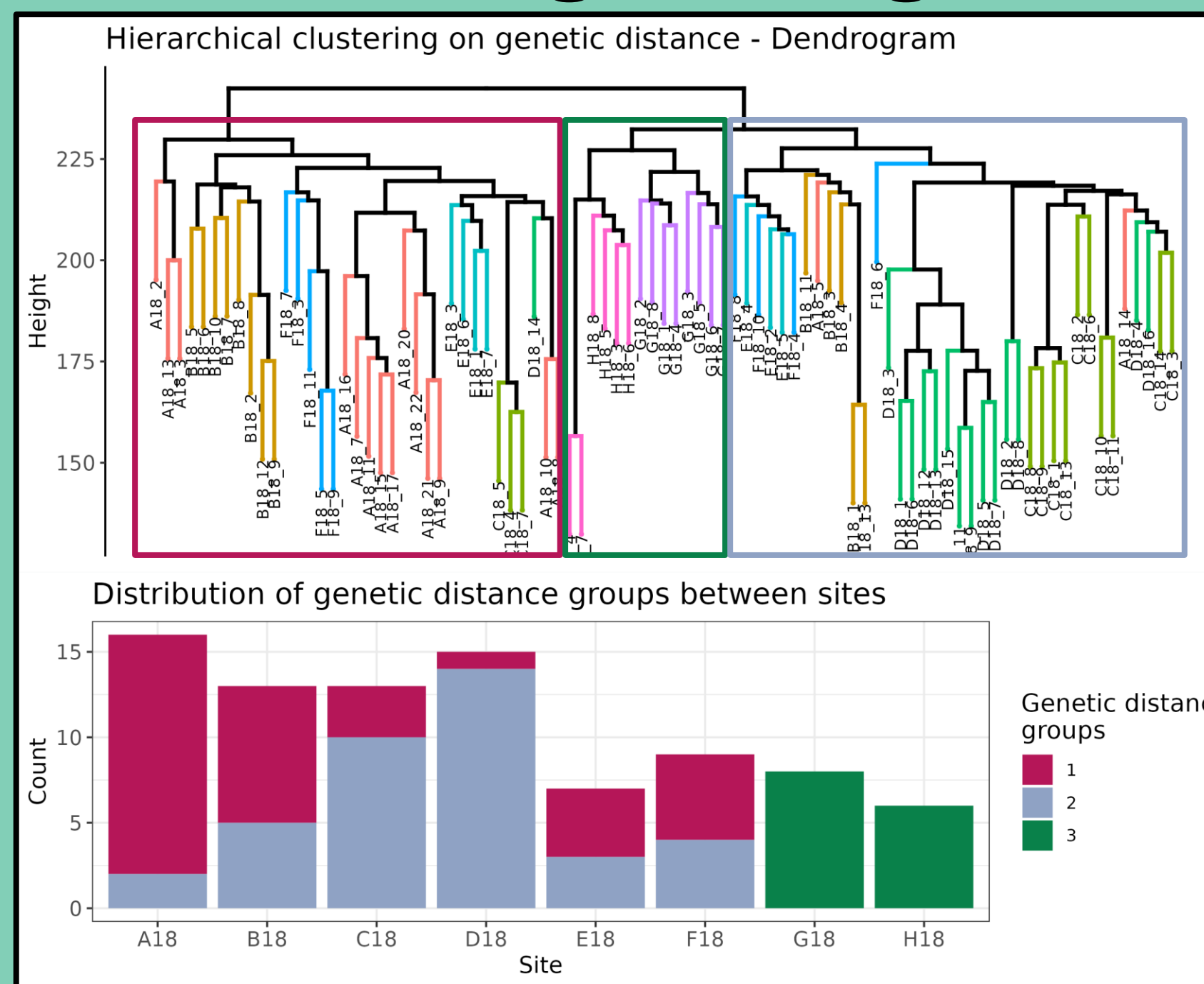
Results
Visualization of results through PCA shows no substantial difference between all three methods: the batch effect appears to have been removed, but so is the variation in the gene expression that was correlated with radiation exposure.



The site: an obscure variable

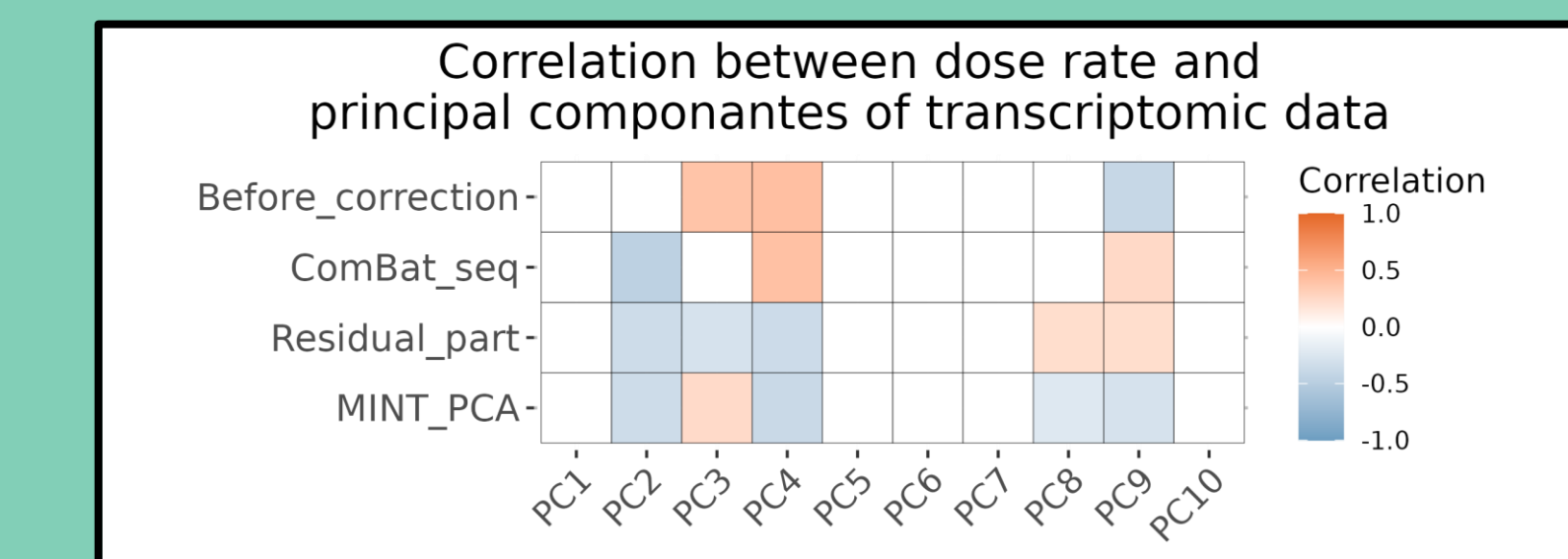
This 8-level categorical variable encapsulates information on site-to-site variations such as local environment and food sources, day of capture, and genetic diversity between sampled populations.

IV. Controlling for the genetic diversity



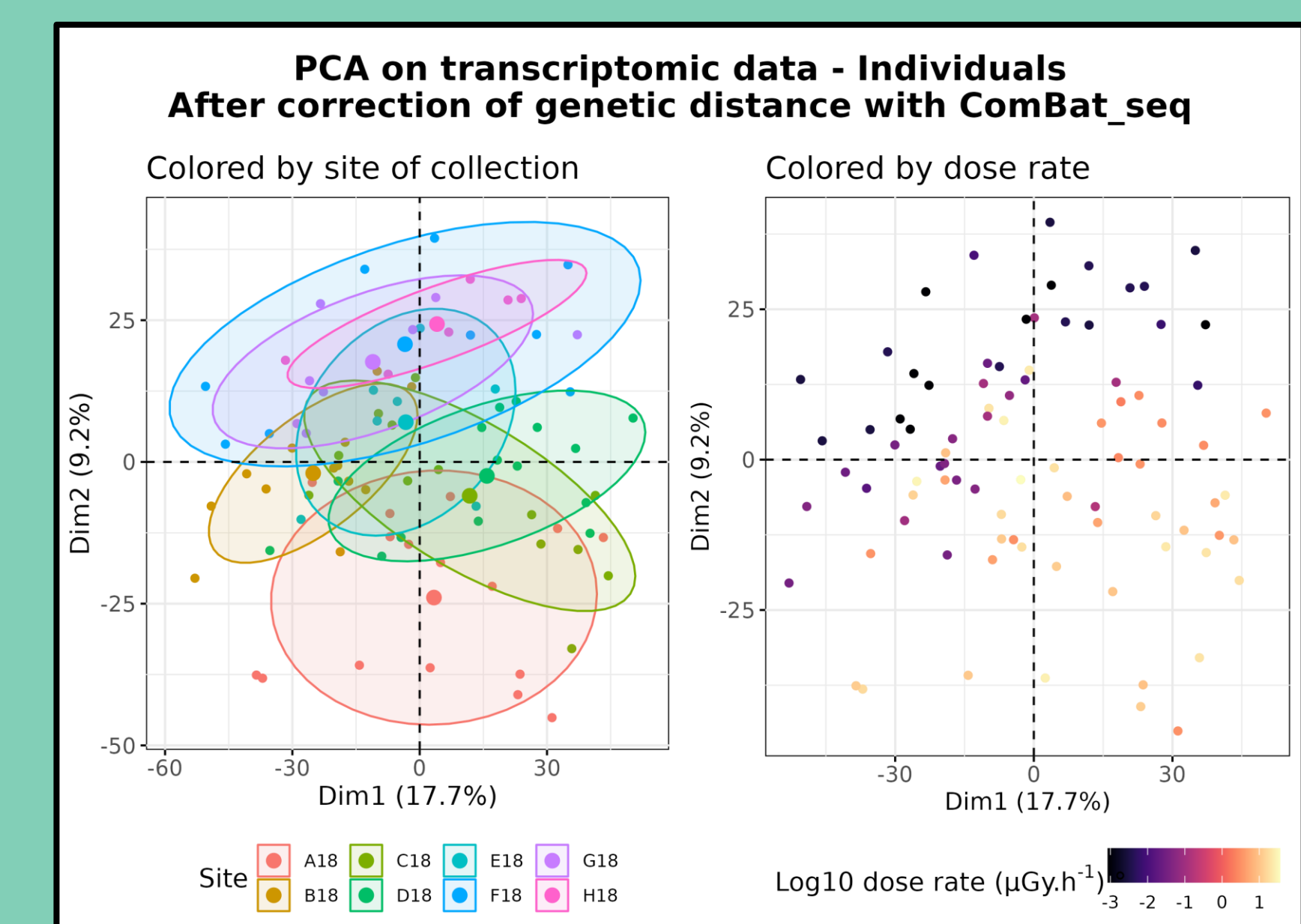
Results

Residualization, ComBat_seq, and MINT PCA yielded qualitatively satisfactory results: monitoring the effect of genetic distance revealed an association between increased dose rate and variation in gene expression.



Treating genetic diversity as a batch effect

To take genetic diversity into account, hierarchical clustering was performed on the inter-individual genetic distance matrix. This hierarchical clustering enabled genetic diversity to be summarized in a 3-level categorical variable. The groups obtained reflect the coherence between genetic structure and geography and are consistent with the migrations observed between sites within the exclusion zone [5]. The methods described previously were applied to the RNA-seq count data to correct/account for the genetic diversity grouping.



V. Conclusion and perspectives

Limits of BE mitigation methods:

- Lack of reliability when applied to a condition-confounded batch.
- ComBat_seq's condition input parameter allows the preservation of condition-related variation but only accepts qualitative variables.
- To reduce the effect of the site, we focused on taking genetic diversity into account as a batch factor, allowing us to extract the variation in expression associated with exposure.

Perspectives include:

- Multi-modal omics dataset integration (RNA-seq, DNA-seq, proteomics).

References

- Burraco, P. *et al.* Assessment of exposure to ionizing radiation in Chernobyl tree frogs (*Hyla orientalis*). *Sci Rep* (2021)
- Car, C. *et al.* Unusual evolution of tree frog populations in the Chernobyl exclusion zone. *Evol Appl* (2022)
- Zhang, Y. *et al.* ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom Bioinform* (2020)
- Rohart, F. *et al.* MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. *BMC Bioinform* (2017)
- Car, C., ..., Goujon, E. *et al.* Population transcriptogenomics highlights impaired metabolism and small population sizes in tree frogs living in the Chernobyl Exclusion zone. *BMC Biol* (Accepted, 2023)