BIOPUCES SEMINAR 12/06/2025

ACCOUNTING FOR CONFOUNDING VARIATION IN MULTI-OMICS DATA INTEGRATION

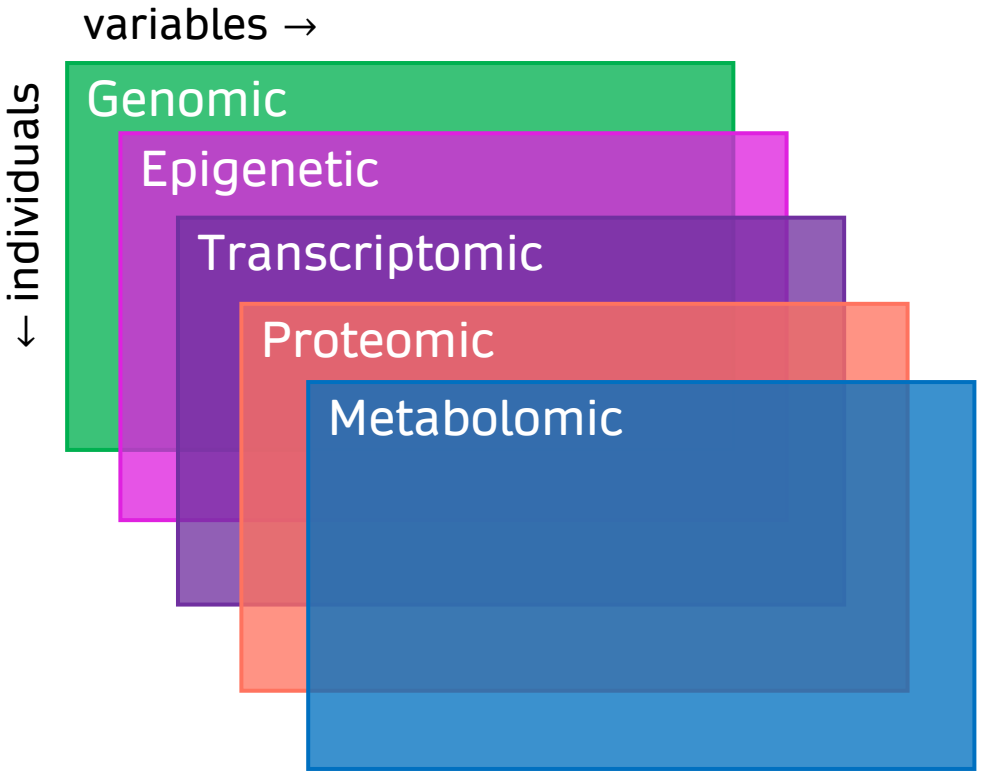
APPLICATION TO THE STUDY OF LOW-DOSE RADIATION EFFECTS IN CHORNOBYL TREE FROGS

Elen Goujon[1,2], Olivier Armant[3], Arthur Tenenhaus[2], Imène Garali[1]

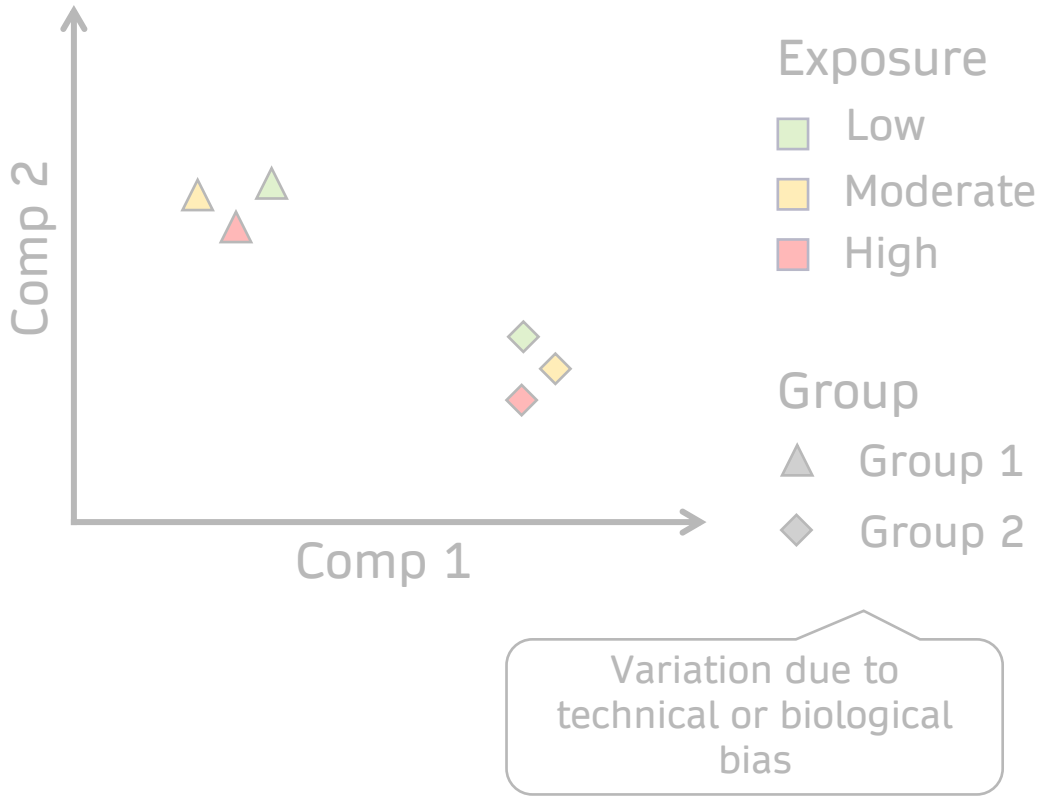
1. ASNR-IRSN, Laboratoire de radiobiologie des expositions accidentelles (LRAcc)
2. CentraleSupélec, CNRS, Université Paris-Saclay, Laboratoire des Signaux et Systèmes (L2S)
3. ASNR-IRSN, Laboratoire de recherche sur les effets des radionucléides sur les écosystèmes (LECO)

METHODOLOGICAL CHALLENGES

How to integrate several omics datasets in a joint analysis?

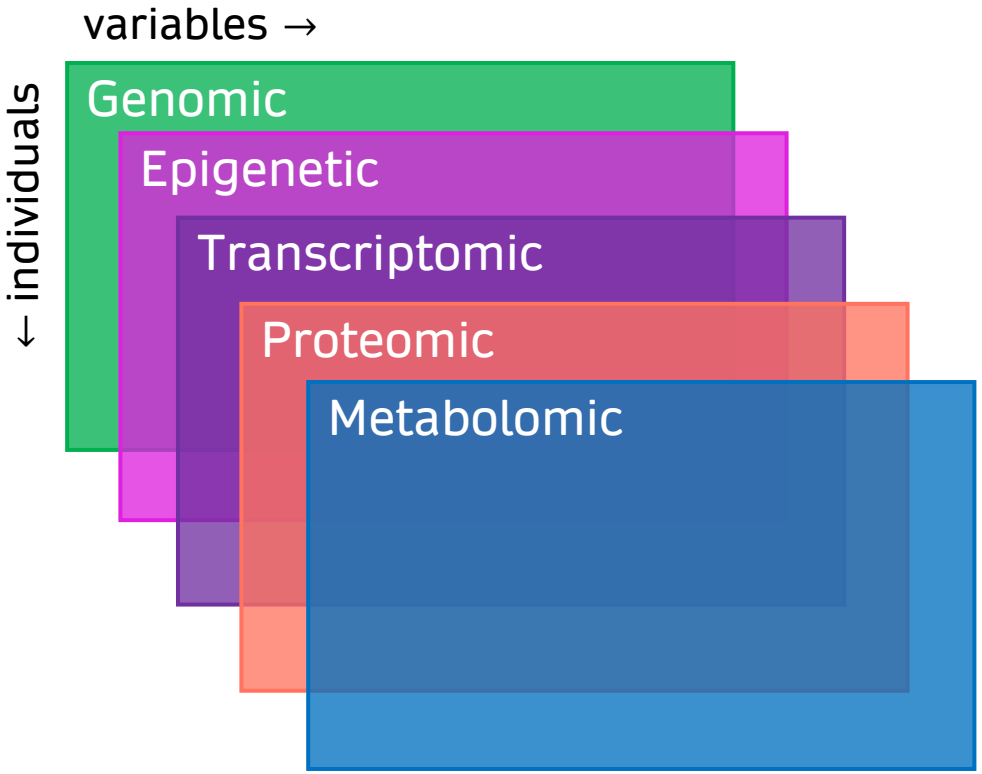


How to manage the effect of undesirable or confounding variables?

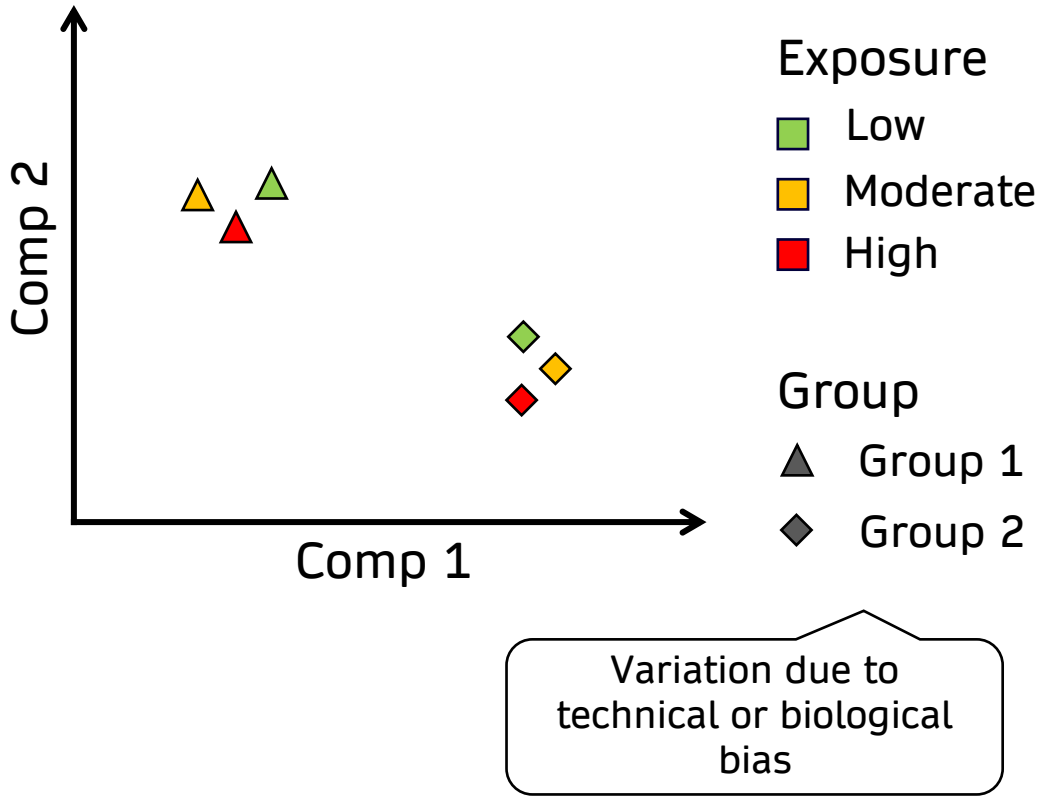


METHODOLOGICAL CHALLENGES

How to integrate several omics datasets in a joint analysis?



How to manage the effect of undesirable or confounding variables?



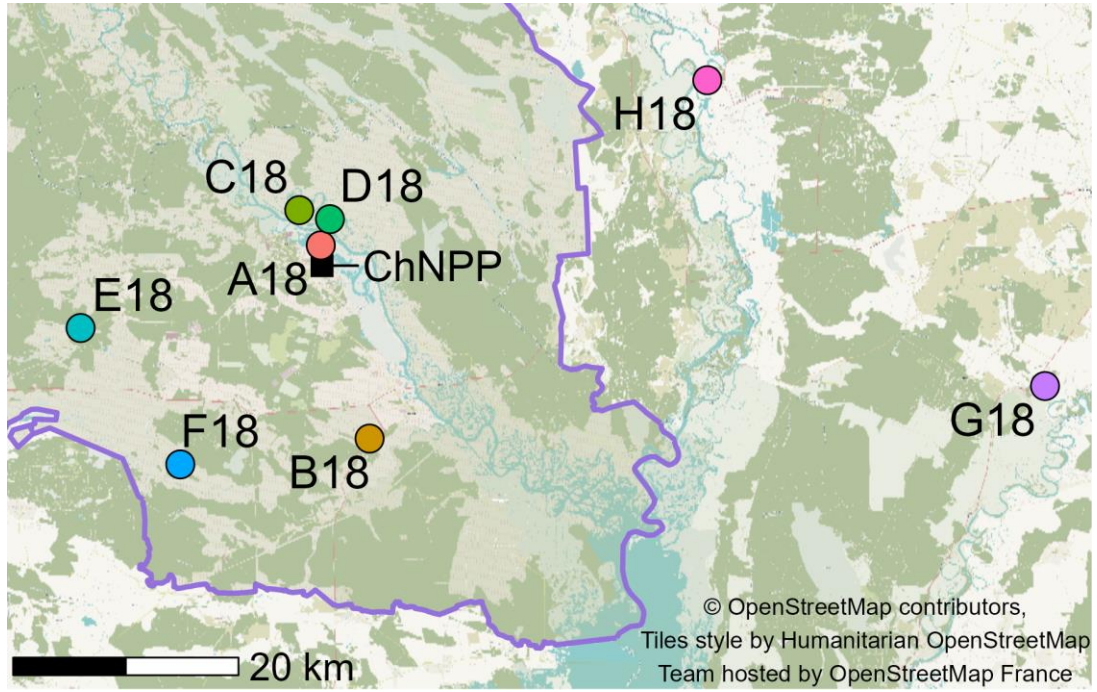
CONTEXT OF THE RADIOPROTECTION STUDY
CHORNOBYL TREE FROGS

Can we identify molecular signatures of chronic exposure to low doses?

2018: sampling of *Hyla orientalis* tree frog populations [Burraco 2021, Car 2022, Car 2023]

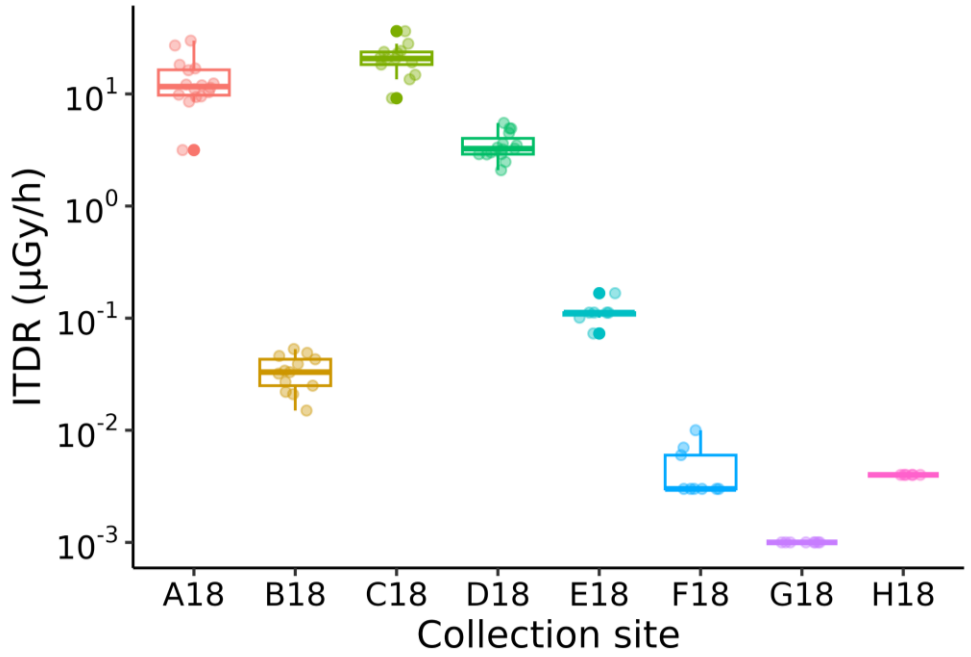
Types of data collected:

- Dosimetry (ITDR = Individual Total Dose Rate, including internal and external contributions from ^{137}Cs and ^{90}Sr)
- Age, phenotype (mass, dimensions)
- Genomics
- Transcriptomics
- Proteomics

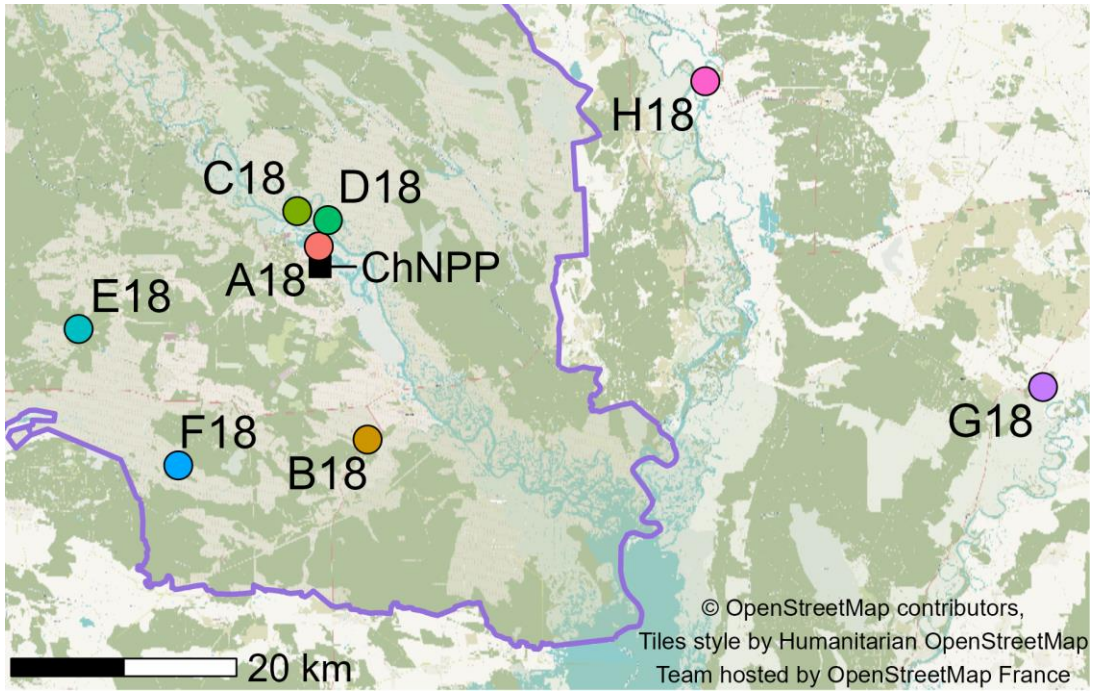


Collection sites inside and outside the Chornobyl Exclusion Zone (CEZ: —, ChNPP = Chornobyl nuclear power plant)

CONTEXT
FROGS' COLLECTION SITES AND RADIATION DOSE RATE

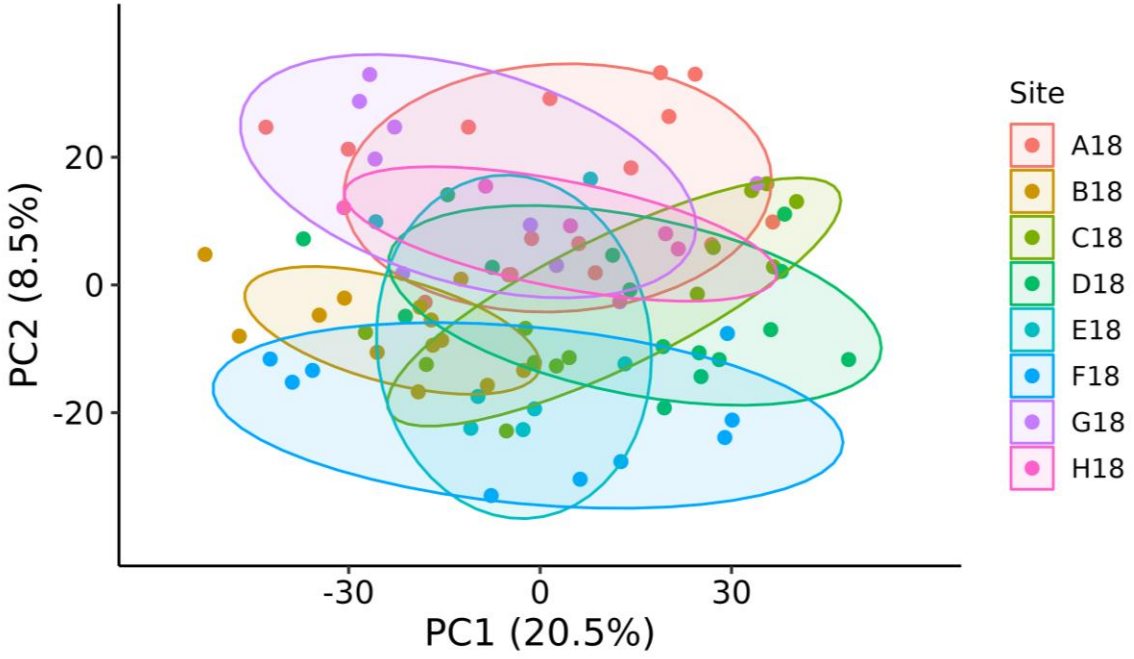


Individual total dose rate (ITDR) distribution among sites

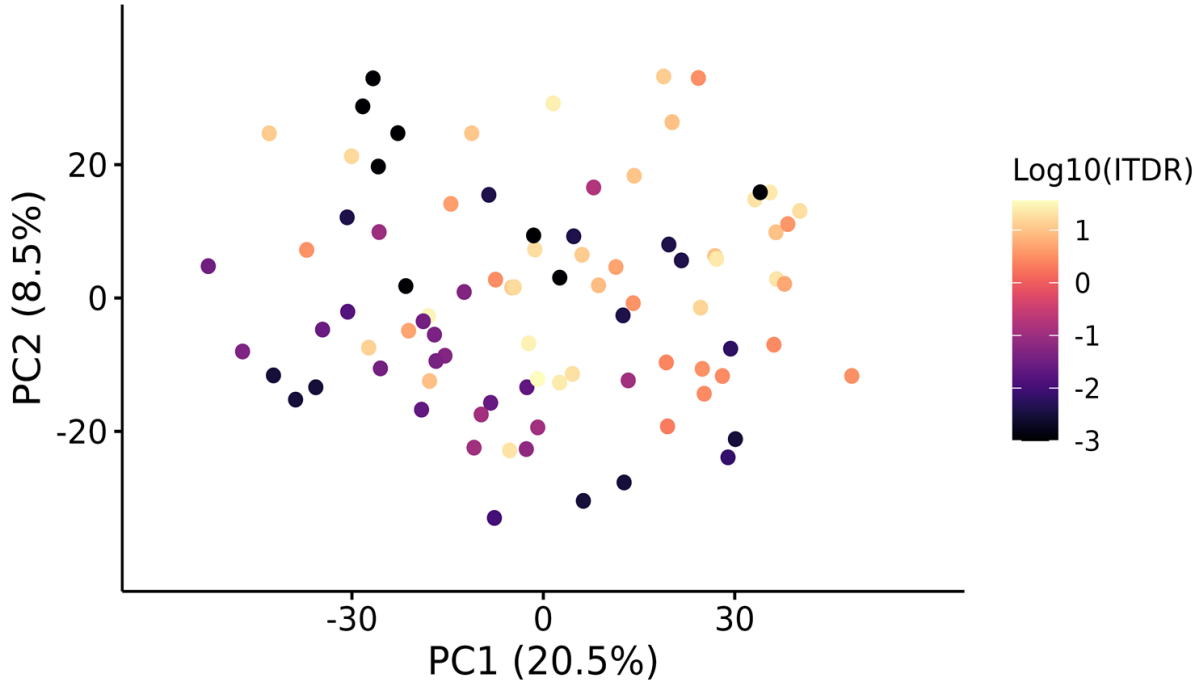


Collection sites inside and outside the Chornobyl Exclusion Zone (CEZ — , ChNPP = Chornobyl nuclear power plant)

CONTEXT
CONFOUNDING EFFECT OF THE SITE ON RNA-SEQ DATA



Individuals in the first factorial plane, colored by site



Individuals in the first factorial plane, colored by dose rate

OBJECTIVES

Advance our understanding of low-dose radiation effects

→ Design an integrative approach for the joint analysis of multi-omics data

→ Handle the confounding effect of the collection site

OUTLINE

HANDLING THE SITE EFFECT IN THE FROGS RNA-SEQ DATA

TOWARDS ACCOUNTING FOR COVARIATES IN MULTI-OMICS ANALYSIS

OUTLINE

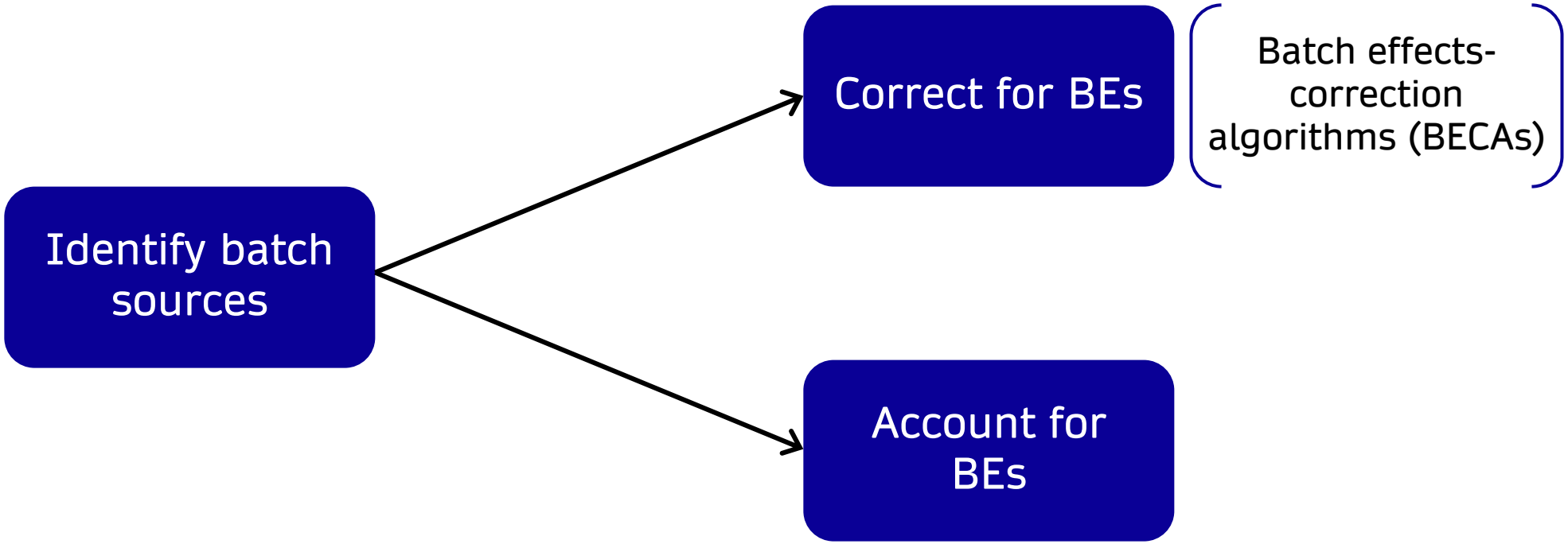
HANDLING THE SITE EFFECT IN THE FROGS RNA-SEQ DATA

TOWARDS ACCOUNTING FOR COVARIATES IN MULTI-OMICS ANALYSIS

HANDLING BATCH EFFECTS (BEs)

Batch effects = biases arising from technical variation [Leek 2010, Goh 2022]

⟶ Can interfere with the analysis of omics data



BATCH EFFECT-CORRECTION ALGORITHMS

Residualization

[Sims 2008, García 2020]
R package `stats` or `limma`

Removes linear relationship between variables and batch

For gene g:

gene expression

batch

biological condition

$$X_g = \hat{\beta}_g^\top z + \hat{\alpha}_g\, y + \varepsilon_g$$

residuals

$$X_g^* = \hat{\alpha}_g y + \varepsilon_g$$

batch-corrected gene expression

BATCH EFFECT-CORRECTION ALGORITHMS

Residualization

[Sims 2008, García 2020]
R package `stats` or `limma`

Removes linear relationship between variables and batch

For gene g:

gene expression

batch

biological condition

$$X_g = \hat{\beta}_g^\top z + \hat{\alpha}_g\, y + \varepsilon_g$$

residuals

$$X_g^* = \hat{\alpha}_g y + \varepsilon_g$$

batch-corrected gene expression

ComBat_seq

[Zhang 2020]
R package `sva`

Corrects data distribution to fit a batch-free distribution

For gene g, sample j, batch i:

$$X_{gij} \sim NB(\mu_{gij}, \phi_{gi})$$

$$\log \mu_{gij} = \alpha_g + \beta_g y_i + \gamma_{gi} + \log N_j$$

average level of expression

biological condition

batch term

library size

BATCH EFFECT-CORRECTION ALGORITHMS

Residualization
[Sims 2008, García 2020]
R package `stats` or `limma`

Removes linear relationship between variables and batch

For gene g:

$$X_g = \hat{\beta}_g^\top z + \hat{\alpha}_g\, y + \varepsilon_g$$

gene expression · batch · biological condition · residuals

$$X_g^* = \hat{\alpha}_g y + \varepsilon_g$$

batch-corrected gene expression

ComBat_seq
[Zhang 2020]
R package `sva`

Corrects data distribution to fit a batch-free distribution

For gene g, sample j, batch i:

$$X_{gij} \sim NB(\mu_{gij}, \phi_{gi})$$

$$\log \mu_{gij} = \alpha_g + \beta_g y_i + \gamma_{gi} + \log N_j$$

average level of expression · biological condition · batch term · library size

Surrogate Variable Analysis (SVA)
[Leek 2007]
R package `sva`

Identifies unobserved sources of expression variation

For gene g:

$$X_g = \hat{\alpha}_g y + \varepsilon_g$$

biological condition · residuals

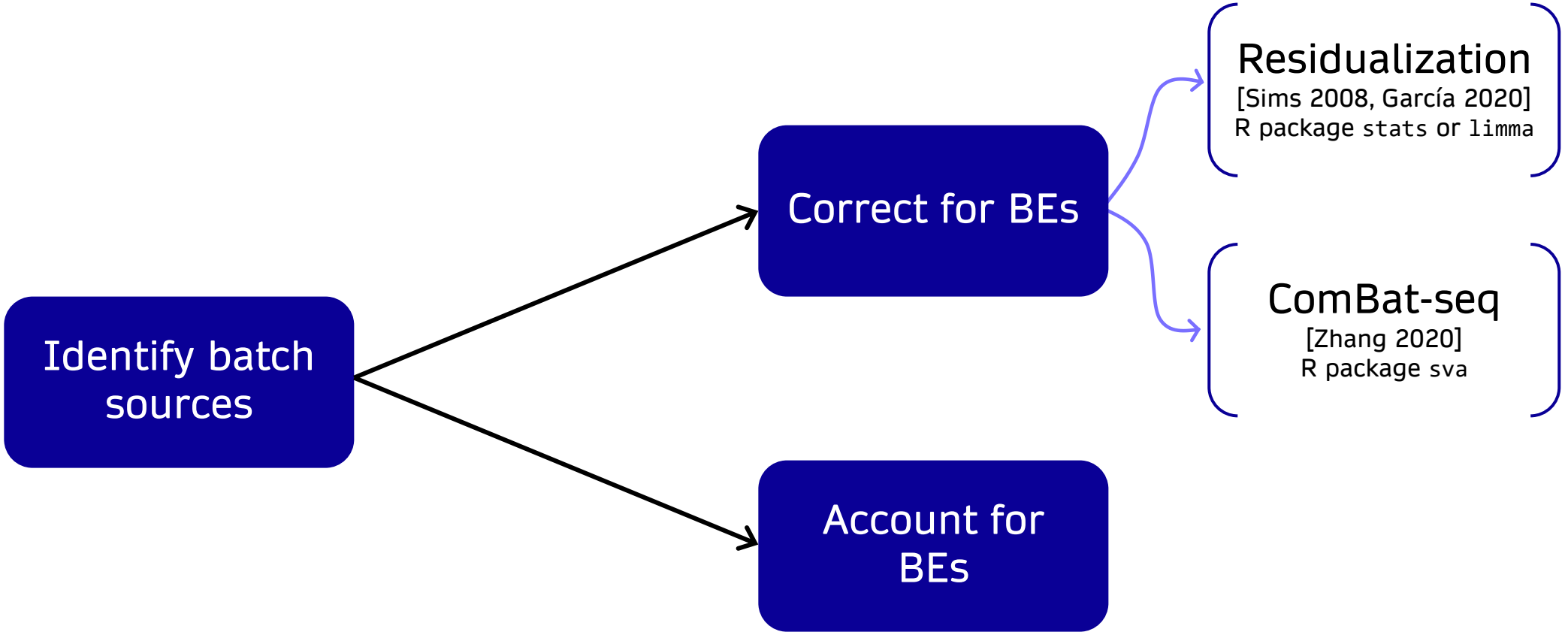
$$X_g^* = \varepsilon_g$$

singular value decomposition

$$X^* = U\Sigma V^\top$$

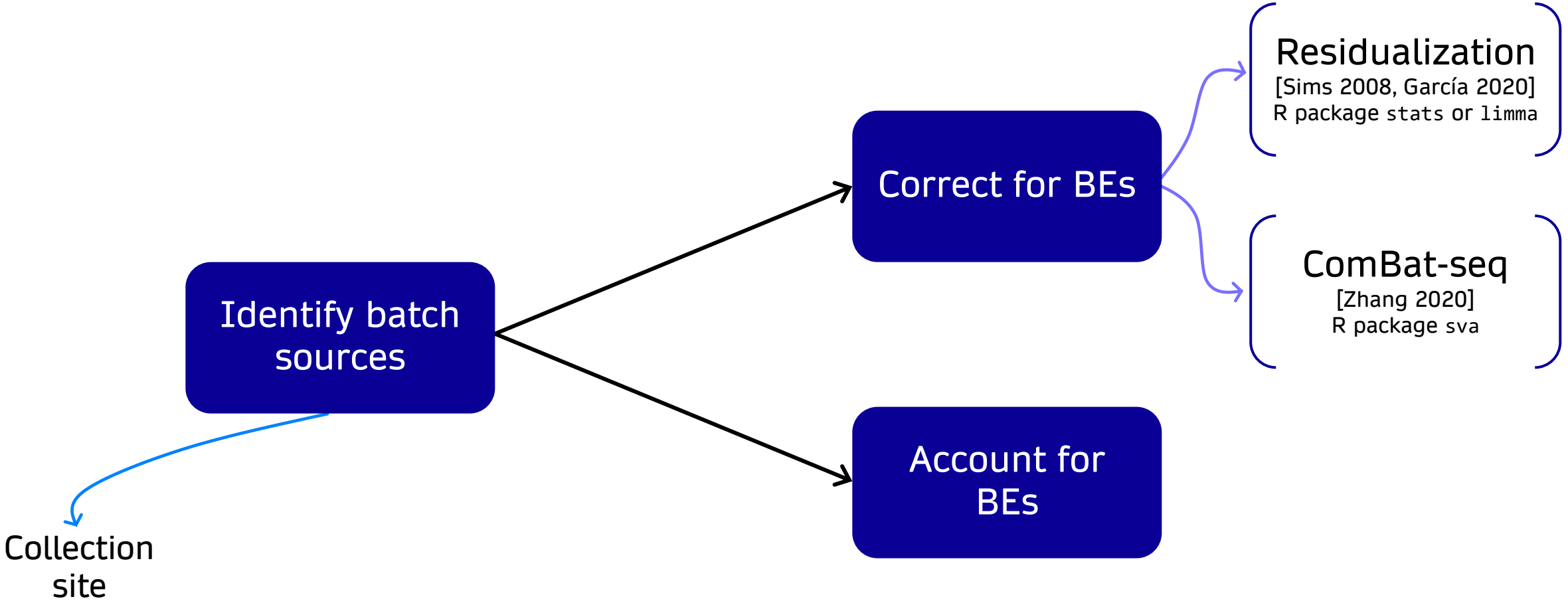
extract surrogate variables

CORRECTION STRATEGIES
RNA-SEQ DATA



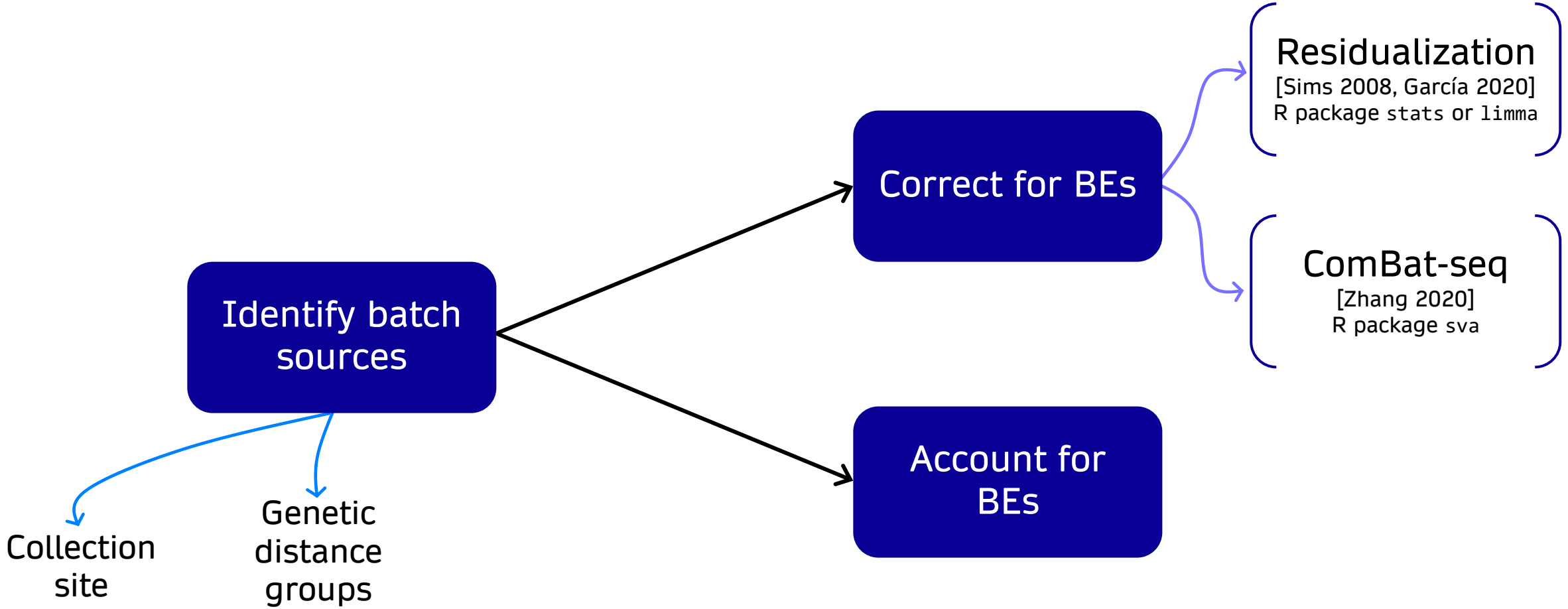
Further information in [Goujon 2024, Car 2023]

CORRECTION STRATEGIES
RNA-SEQ DATA



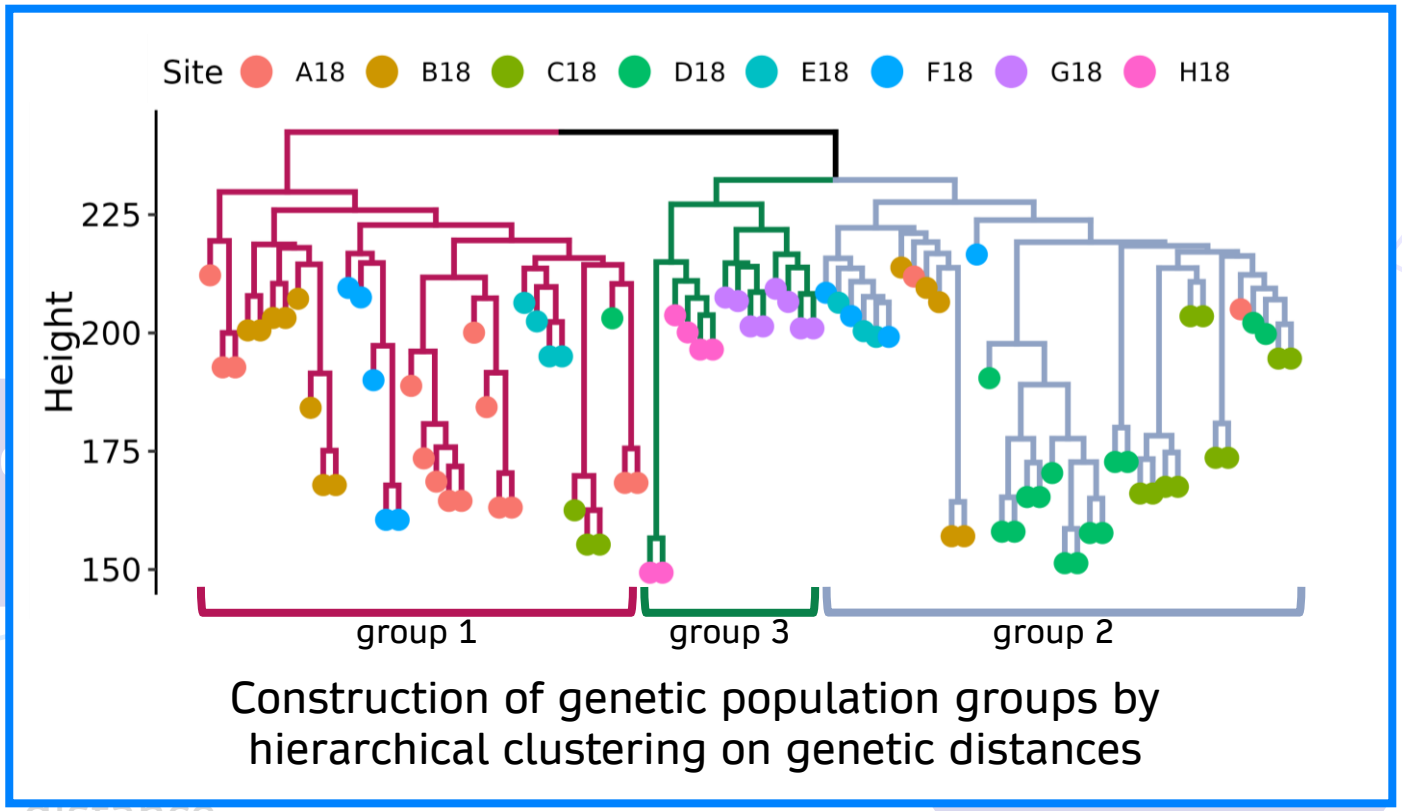
Further information in [Goujon 2024, Car 2023]

CORRECTION STRATEGIES
RNA-SEQ DATA



Further information in [Goujon 2024, Car 2023]

CORRECTION STRATEGIES
RNA-SEQ DATA



Construction of genetic population groups by hierarchical clustering on genetic distances

Residualization
[Sims 2008, García 2020]
R package stats or limma

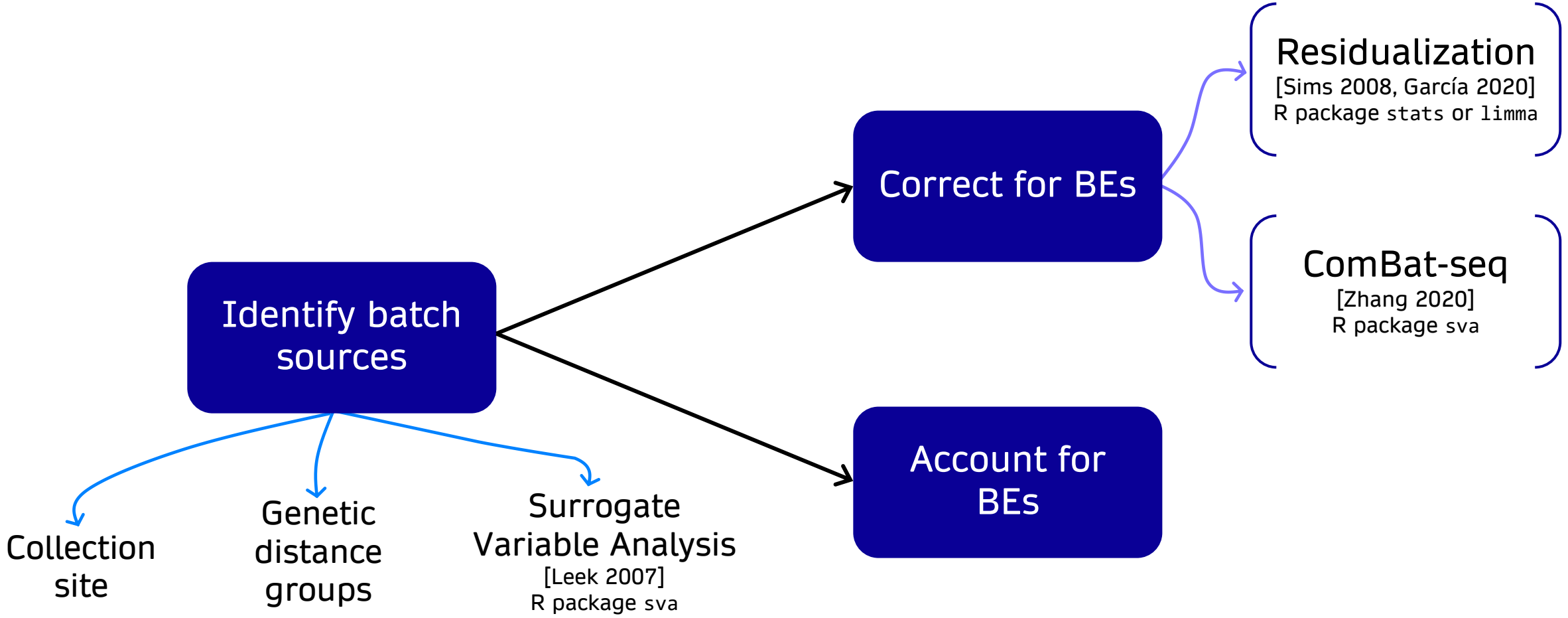
ComBat-seq
[Zhang 2020]
R package sva

Collection site

distance groups

Further information in [Goujon 2024, Car 2023]

CORRECTION STRATEGIES
RNA-SEQ DATA



Further information in [Goujon 2024, Car 2023]

METHODS

Batch effect assessment:

- principal components analysis (PCA) → visualization
  → quantification

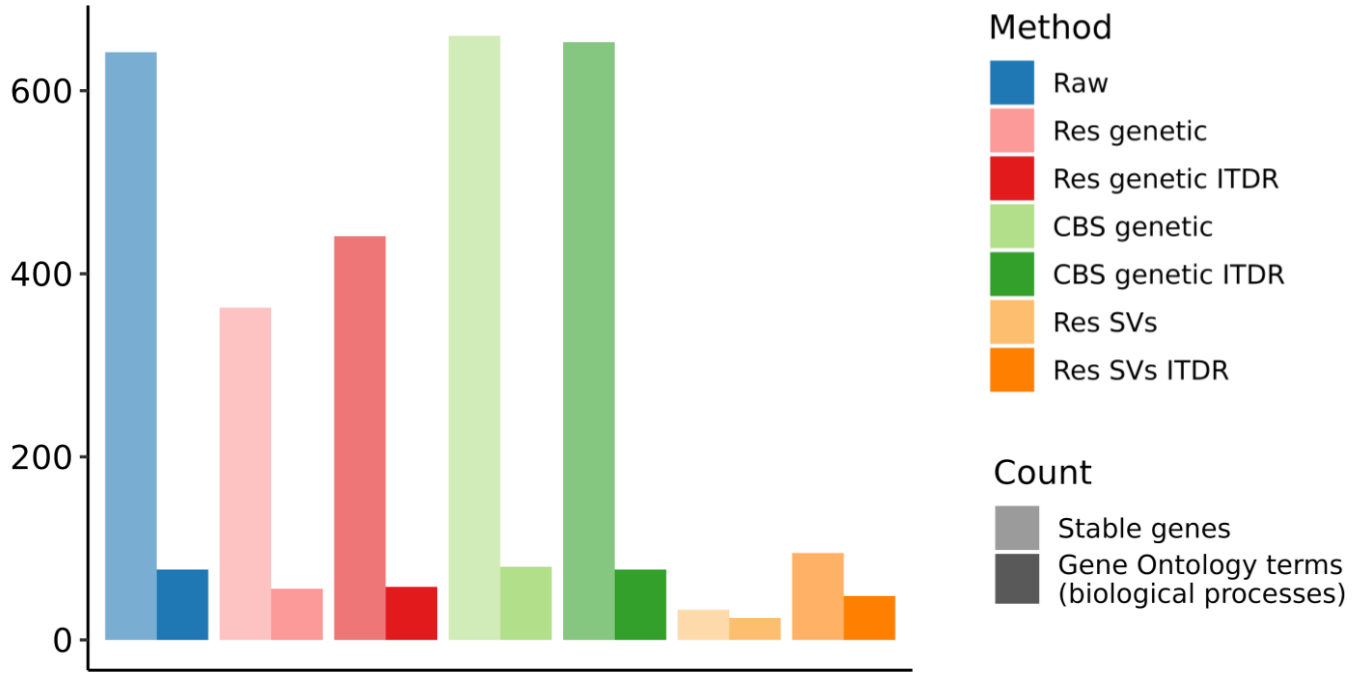
Batch effect correction:

- BECAs: residualization, ComBat-seq, SVA
  - ➢ both with and without preservation of radiocontamination effects

Performance evaluation:

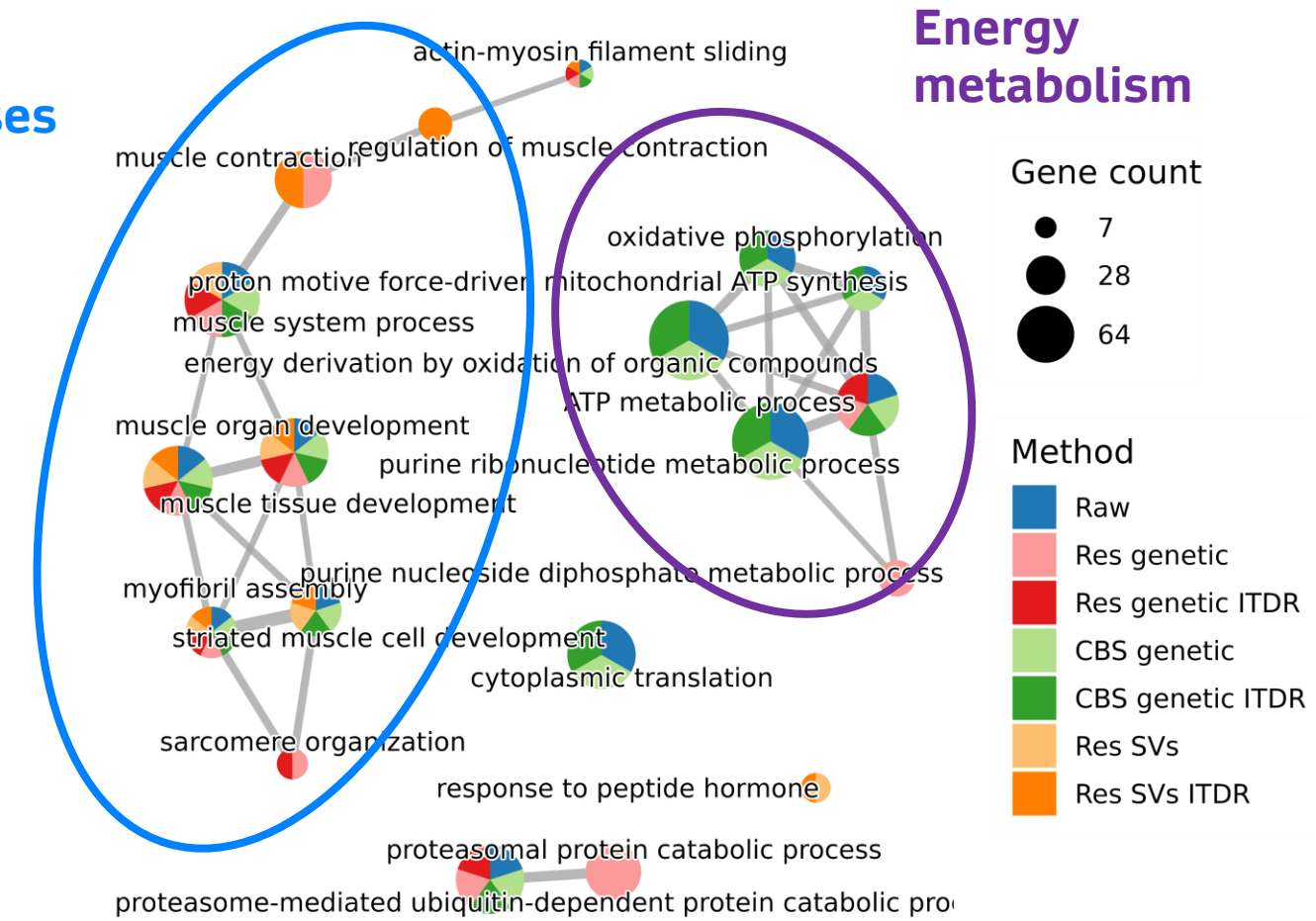
- visualization and summary statistics
- biological interpretation
  - ➢ sparse PCA followed by functional enrichment analysis

RESULTS
BIOLOGICAL INFORMATION IN CORRECTED DATASETS



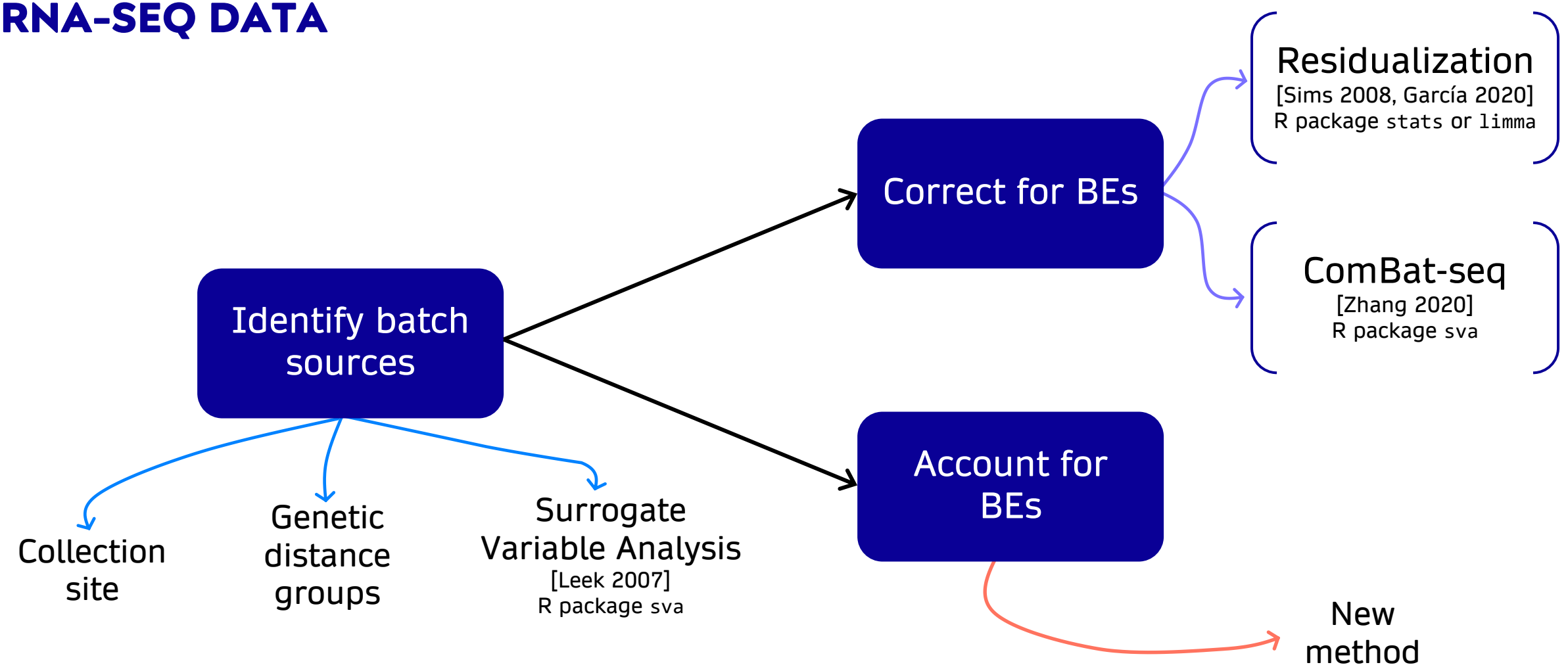
Number of stable genes selected in PCs 1 and 2, and enriched GO terms

RESULTS
ENRICHED GENE ONTOLOGY TERMS



Network of enriched GO terms after the different correction strategies

CORRECTION STRATEGIES
RNA-SEQ DATA



Further information in [Goujon 2024, Car 2023]

OUTLINE

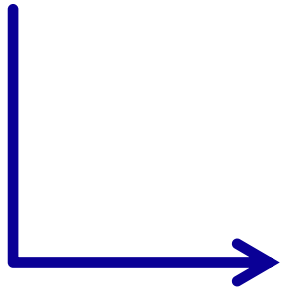
HANDLING THE SITE EFFECT IN THE FROGS RNA-SEQ DATA

TOWARDS ACCOUNTING FOR COVARIATES IN MULTI-OMICS ANALYSIS

TOWARDS ACCOUNTING FOR COVARIATES IN MULTI-OMICS ANALYSIS

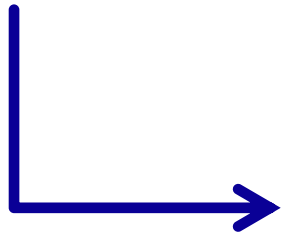
partial correlation coefficient

$r(x_1, x_2|z)$



partial CCA [Rao 1969]

$\mathbf{X}_1, \mathbf{X}_2 \mid \mathbf{Z}$



generalization to multi-block component analysis

AC-PCA

SIMULTANEOUS DIMENSION REDUCTION AND ADJUSTMENT FOR CONFOUNDING [LIN 2016]

⟶ Penalty scheme used to take confounding factors into account

- dataset $\mathbf{X} \in \mathcal{M}_{n,p}$
- confounders matrix $\mathbf{Y} \in \mathcal{M}_{n,l}$
- kernel matrix $\mathbf{K} = \mathbf{Y}\mathbf{Y}^\top$
- parameter $\lambda > 0$

$$\max_{\mathbf{v} \in \mathbb{R}^p} \mathbf{v}^\top \mathbf{X}^\top \mathbf{X} \mathbf{v} - \lambda \mathbf{v}^\top \mathbf{X}^\top \mathbf{Y}\mathbf{Y}^\top \mathbf{X} \mathbf{v}$$
$$\text{subject to } \|\mathbf{v}\|_2^2 \leq 1$$

AC-PCA

SIMULTANEOUS DIMENSION REDUCTION AND ADJUSTMENT FOR CONFOUNDING [LIN 2016]

⟶ Penalty scheme used to take confounding factors into account

- dataset $\mathbf{X} \in \mathcal{M}_{n,p}$
- confounders matrix $\mathbf{Y} \in \mathcal{M}_{n,l}$
- kernel matrix $\mathbf{K} = \mathbf{Y}\mathbf{Y}^\top$
- parameter $\lambda > 0$

$$\max_{\mathbf{v} \in \mathbb{R}^p} \mathbf{v}^\top \mathbf{X}^\top \mathbf{X} \mathbf{v} \; - \lambda \mathbf{v}^\top \mathbf{X}^\top \mathbf{Y}\mathbf{Y}^\top \mathbf{X} \mathbf{v}$$
$$\text{subject to } \|\mathbf{v}\|_2^2 \; \leq 1$$

AC-PCA

SIMULTANEOUS DIMENSION REDUCTION AND ADJUSTMENT FOR CONFOUNDING [LIN 2016]

⟶ Penalty scheme used to take confounding factors into account

- dataset $\mathbf{X} \in \mathcal{M}_{n,p}$
- confounders matrix $\mathbf{Y} \in \mathcal{M}_{n,l}$
- kernel matrix $\mathbf{K} = \mathbf{Y}\mathbf{Y}^\top$
- parameter $\lambda > 0$

$$\max_{\mathbf{v} \in \mathbb{R}^p} \mathbf{v}^\top \mathbf{X}^\top \mathbf{X}\mathbf{v} - \lambda \mathbf{v}^\top \mathbf{X}^\top \mathbf{Y}\mathbf{Y}^\top \mathbf{X}\mathbf{v}$$
$$\text{subject to } \|\mathbf{v}\|_2^2 \leq 1$$

$\mathbf{Y}^\top \mathbf{X}\mathbf{v} = \text{cov}(\mathbf{Y}, \mathbf{X}\mathbf{v})$

AC-PCA

SIMULTANEOUS DIMENSION REDUCTION AND ADJUSTMENT FOR CONFOUNDING [LIN 2016]

⟶ Penalty scheme used to take confounding factors into account

- dataset $\mathbf{X} \in \mathcal{M}_{n,p}$
- confounders matrix $\mathbf{Y} \in \mathcal{M}_{n,l}$
- kernel matrix $\mathbf{K} = \mathbf{Y}\mathbf{Y}^\top$
- parameter $\lambda > 0$

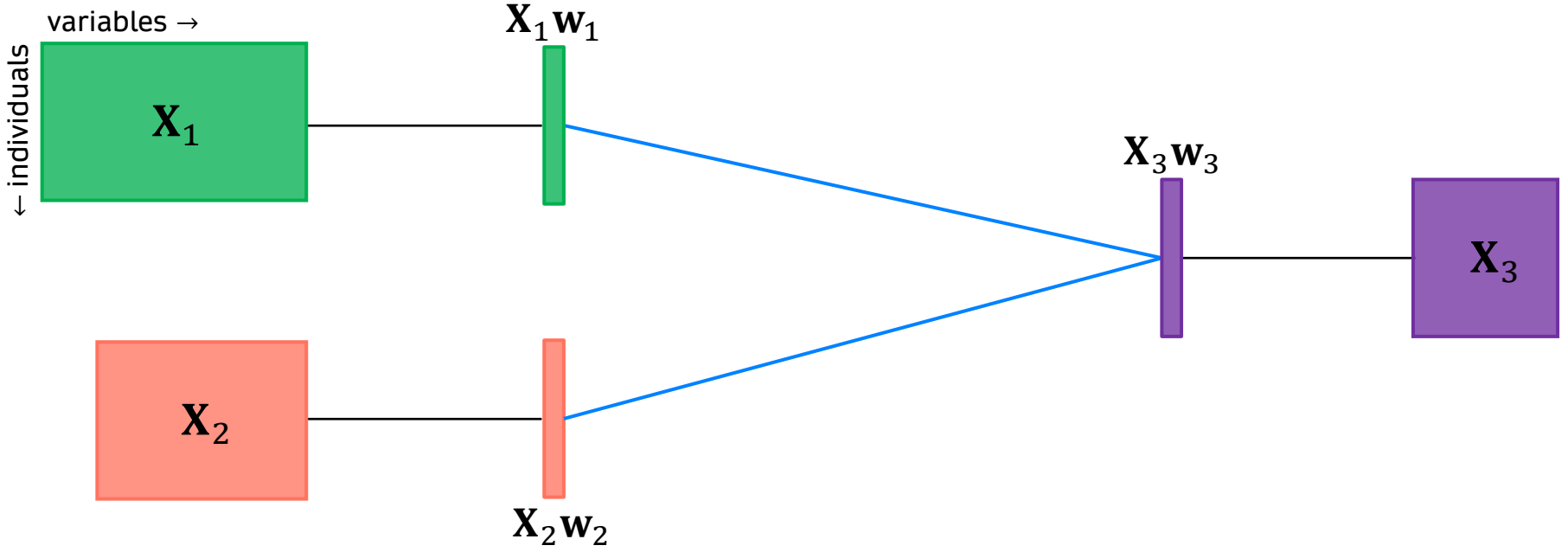
or another positive definite matrix

$$\max_{\mathbf{v} \in \mathbb{R}^p} \mathbf{v}^\top \mathbf{X}^\top \mathbf{X}\mathbf{v} - \lambda \mathbf{v}^\top \mathbf{X}^\top \mathbf{K} \mathbf{X}\mathbf{v}$$
$$\text{subject to } \|\mathbf{v}\|_2^2 \leq 1$$

MULTI-BLOCK RGCCA FRAMEWORK

REGULARIZED GENERALIZED CANONICAL CORRELATION ANALYSIS [TENENHAUS 2014, 2017]

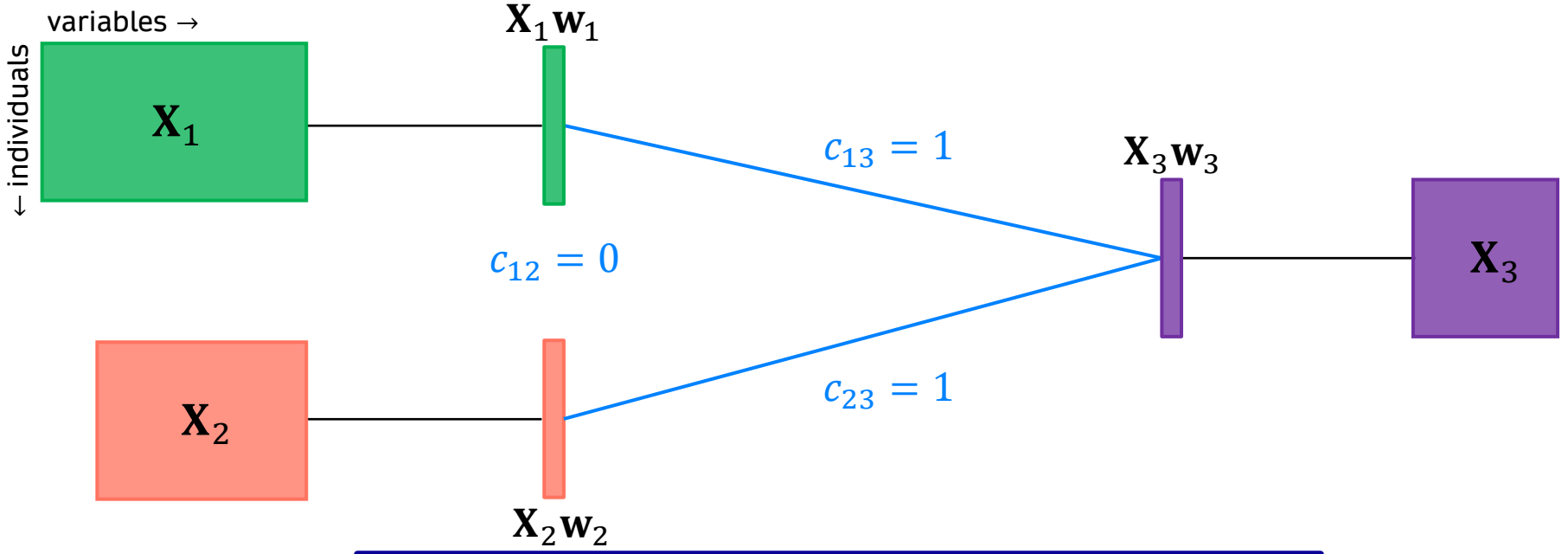
⟶ Statistical framework for multi-omics integration



MULTI-BLOCK RGCCA FRAMEWORK

REGULARIZED GENERALIZED CANONICAL CORRELATION ANALYSIS [TENENHAUS 2014, 2017]

⟶ Statistical framework for multi-omics integration



$$\max_{\mathbf{w}_1,\ldots,\mathbf{w}_J} \sum_{j,k=1}^{J} c_{jk}\, g\left(\mathrm{cov}\left(\mathbf{X}_j\mathbf{w}_j, \mathbf{X}_k\mathbf{w}_k\right)\right)$$

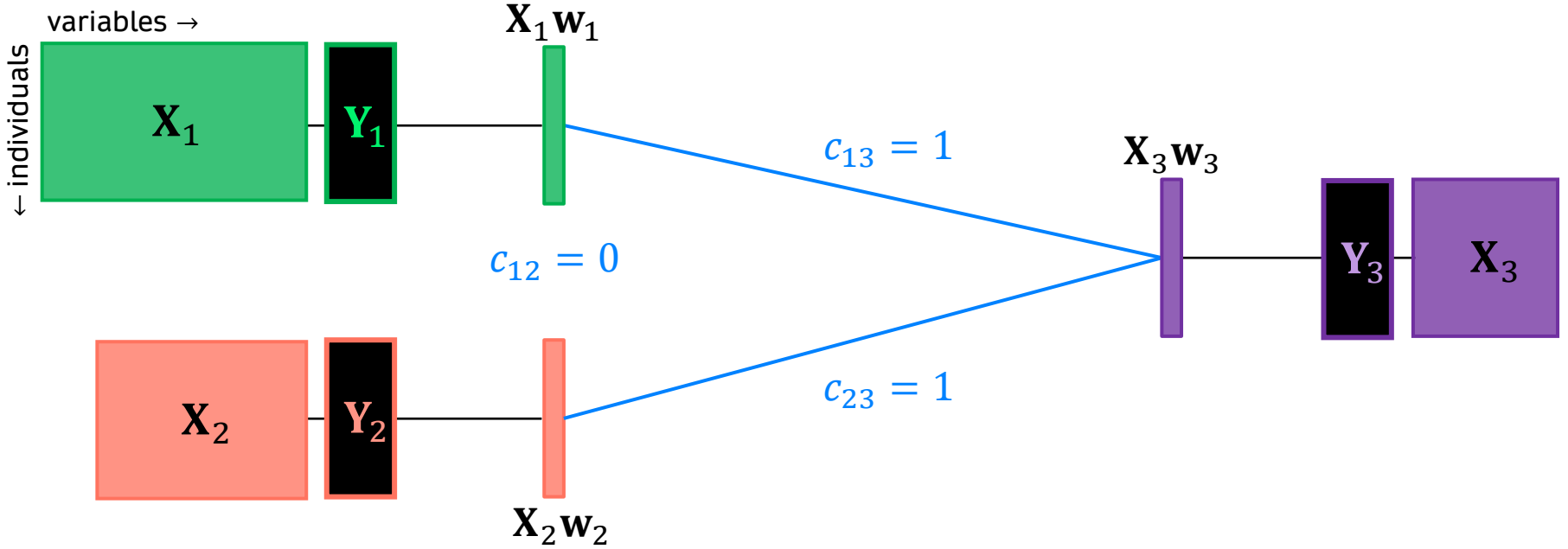
under constraints on $\mathbf{w}_j, j = 1, \ldots, J$

- $c_{jk} = 1$ if $\mathbf{X}_j \leftrightarrow \mathbf{X}_k$, else 0
- g = convex continuous function

MULTI-BLOCK RGCCA FRAMEWORK

REGULARIZED GENERALIZED CANONICAL CORRELATION ANALYSIS [TENENHAUS 2014, 2017]

⟶ Statistical framework for multi-omics integration



$$\max_{\mathbf{w}_1,\ldots,\mathbf{w}_J} \sum_{j,k=1}^{J} c_{jk}\, g\left(\mathrm{cov}\left(\mathbf{X}_j\mathbf{w}_j, \mathbf{X}_k\mathbf{w}_k\right)\right)$$

under constraints on $\mathbf{w}_j, j = 1, \ldots, J$

- $c_{jk} = 1$ if $\mathbf{X}_j \leftrightarrow \mathbf{X}_k$, else 0
- g = convex continuous function

EXTENSION TO AC-RGCCA

⟶ Include the constraint in the multi-block RGCCA framework

penalty parameter

kernel matrix built from the confounders

$$\max_{w_1,\dots,w_J} \sum_{j,k=1}^{J} c_{jk}\, g\left(\mathrm{cov}\left(\mathbf{X}_j\mathbf{w}_j, \mathbf{X}_k\mathbf{w}_k\right)\right) - \sum_{l=1}^{J} \frac{\lambda_l}{n} \mathbf{w}_l^\top \mathbf{X}_l^\top \mathbf{K}_l \mathbf{X}_l \mathbf{w}_l$$

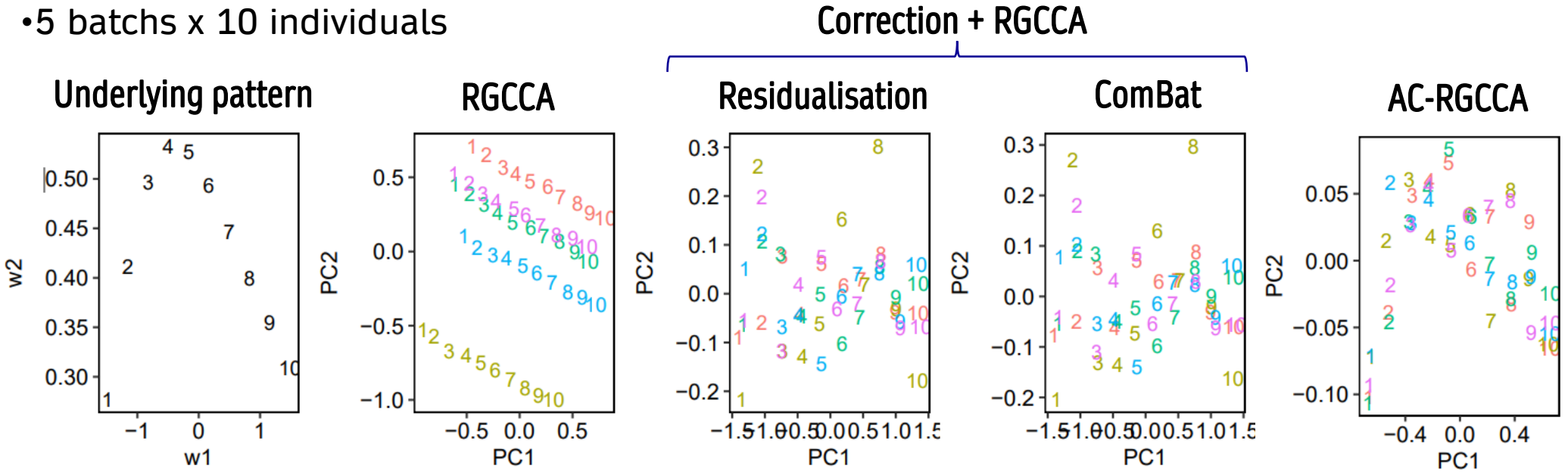
$$\text{under constraints on } \mathbf{w}_j = 1\,, j = 1, \dots, J$$

Iterative algorithm based on block relaxation and Lagragian method
Implementation in R on the *GitHub* of package RGCCA

RESULTS ON SIMULATED DATA

- Method validation
  - Reproduction of AC-PCA's results

- Evaluation of results in multi-block simulations
  - Simulation scenario:
    - 3 blocks
    - 5 batchs x 10 individuals



CONCLUSION & PERSPECTIVES

AXE 1: HANDLING THE SITE EFFECT IN THE FROGS RNA-SEQ DATA

AXE 2: INTEGRATED APPROACHES TO ADJUST FOR CONFOUNDERS

- Multi-group RGCCA [Tenenhaus 2014] (https://github.com/rgcca-factory/RGCCA/tree/multigroup)
- AC-RGCCA

Current work:

- Application to the Chornobyl tree frogs study
- AC-RGCCA: finish writing the article
- Thesis manuscript

VALORIZATIONS

PUBLICATION

BMC BIOLOGY, 2023

“Population transcriptogenomics highlights impaired metabolism and small population sizes in tree frogs living in the Chernobyl Exclusion Zone”
Clément Car, André Gilles, **Elen Goujon**, *et al*.

CONFERENCE PROCEEDINGS PAPER AND ORAL COMMUNICATION

CMSB 2024 (PRÉSENTATION ORALE)

CMSB 2024

“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, and Imène Garali

POSTERS

ICRER 2024

“Leveraging multi-omics data integration in the study of Chornobyl tree frogs”
Elen Goujon, Olivier Armant, Mélodie Noyau, Sandrine Frelon, Luc Camoin, Jean-Marc Bonzom, Arthur Tenenhaus, and Imène Garali

6th International Conference on Radioecology & Environmental Radioactivity
MARSEILLE, FRANCE
24TH–29TH NOVEMBER 2024

JOBIM 2023 (POSTER ET FLASH-TALK)

“Handling confounding factors in analyzing the transcriptomic data from Chornobyl tree frogs”
Elen Goujon, Olivier Armant, Jean-Marc Bonzom, Arthur Tenenhaus, and Imène Garali

JOBIM 2023
27>30 JUIN
Édition multisite

REFERENCES

Burraco, P. *et al*. Assessment Of Exposure To Ionizing Radiation In Chernobyl Tree Frogs (Hyla Orientalis). Scientific Reports 11, 20509 (2021)

Car, C., *et al*. Unusual Evolution Of Tree Frog Populations In The Chernobyl Exclusion Zone. Evolutionary Applications 15(2), 203–219 (2022)

Car, C., *et al*. Population Transcriptogenomics Highlights Impaired Metabolism And Small Population Sizes In Tree Frogs Living In The Chernobyl Exclusion Zone. BMC Biology 21(1), 164 (2023)

Leek, J.T. *et al*. Tackling The Widespread And Critical Impact Of Batch Effects In High-throughput Data. Nature Reviews Genetics 11(10), 733–739 (2010)

Goh, W.W.B., *et al*. Are Batch Effects Still Relevant In The Age Of Big Data? Trends In Biotechnology 40(9), 1029–1040 (2022)

Sims, A.H., *et al*. 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)

García, C.B., *et al*. Residualization: Justification, Properties And Application. Journal Of Applied Statistics 47(11), 1990–2010 (2020)

Zhang, Y., *et al.* Combat-seq: Batch Effect Adjustment For Rna-seq Count Data. NAR Genomics And Bioinformatics 2(3), Lqaa078 (2020)

Goujon, E. *et al.* Batch Effect Correction In A Confounded Scenario: A Case Study On Gene Expression Of Chornobyl Tree Frogs. CMSB Proceedings In Lecture Notes In Computer Science, 14971. (2024)

Lin, Z., *et al.* Simultaneous Dimension Reduction And Adjustment For Confounding Variation. Proceedings Of The National Academy Of Sciences Of The United States Of America, 113(51), 14662–14667. (2016)

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)

THANK YOU

LRAcc
Imène Garali

LRTOX
Mélodie Noyau

LECO
Olivier Armant
Jean-Marc Bonzom
Clément Car
Sandrine Frelon

L2S
Arthur Tenenhaus
Fabien Girka
Laurent Le Brusquet