# Higher-order correction of persistent batch effects in correlation networks

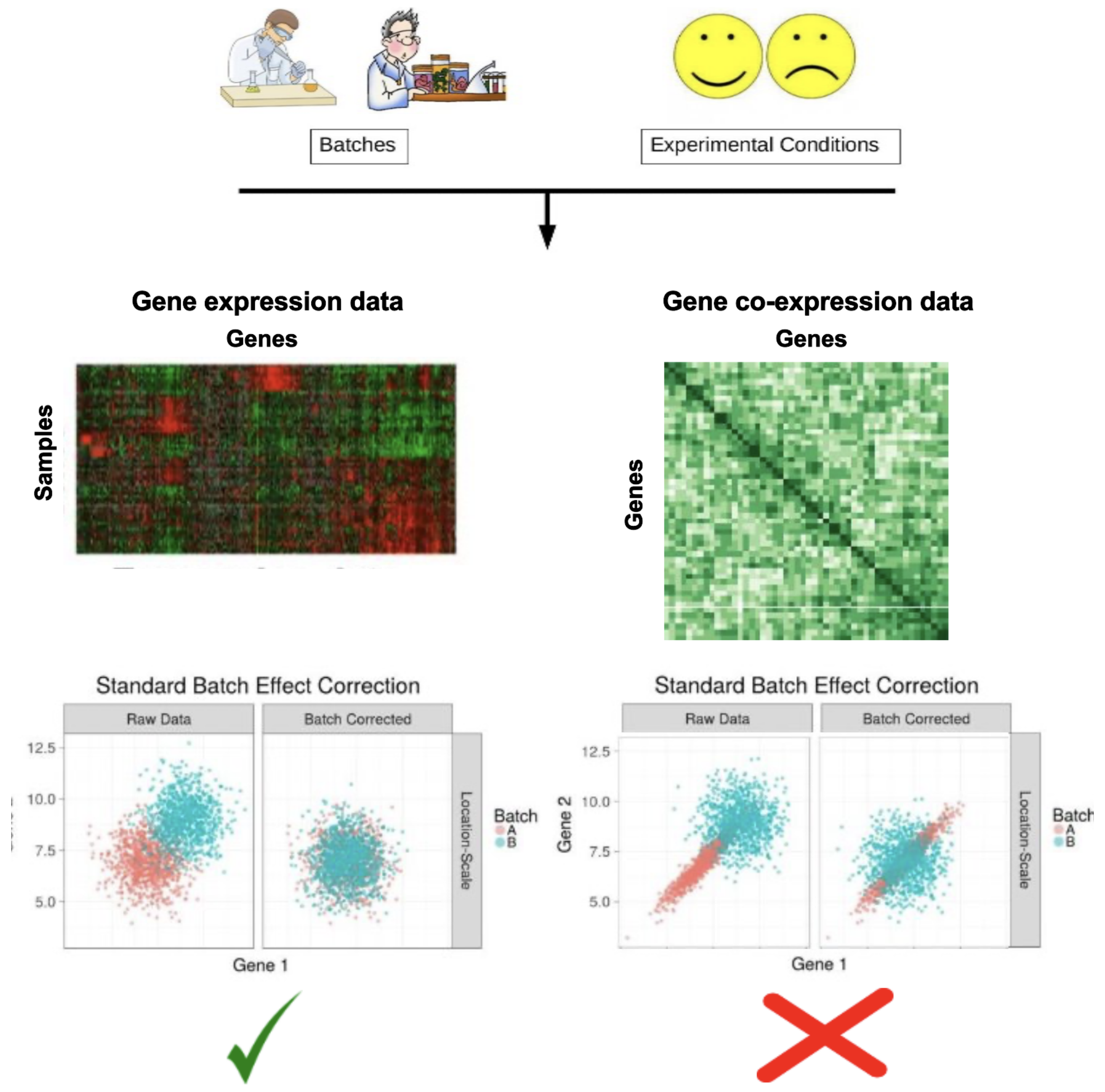
Soel Micheletti[1*], Daniel Schlauch[1,2,3*], John Quackenbush[1, 3, 4†], Marouen Ben Guebila[1†]

[1] Harvard T.H. Chan School of Public Health (Boston, MA), [2] Genospace LLC, (Boston, MA), [3] Dana-Farber Cancer Institute (Boston, MA), [4]Channing Division of Network Medicine,(Boston, MA)

## Objectives

**Gene expression** data is typically collected or processed in different groups or **batches**.

As a community, computational biologists have long recognized that proper analysis of this data relies on **correcting** for systematic differences in experimental settings, usually referred to as **”batch effects”**.



Current batch correction methods do not account for differences in co-expression between batches. This can lead to **biased results** in important applications, including:

- (differential) gene co-expression analysis,
- gene regulatory network (GRN) inference.

**Contributions:**

➜ We show the existence of **residual batch effects** in gene co-expression data **after standard batch correction**.

➜ We present a **new batch correction method** to effectively **identify** and **remove** these spurious residual differences.

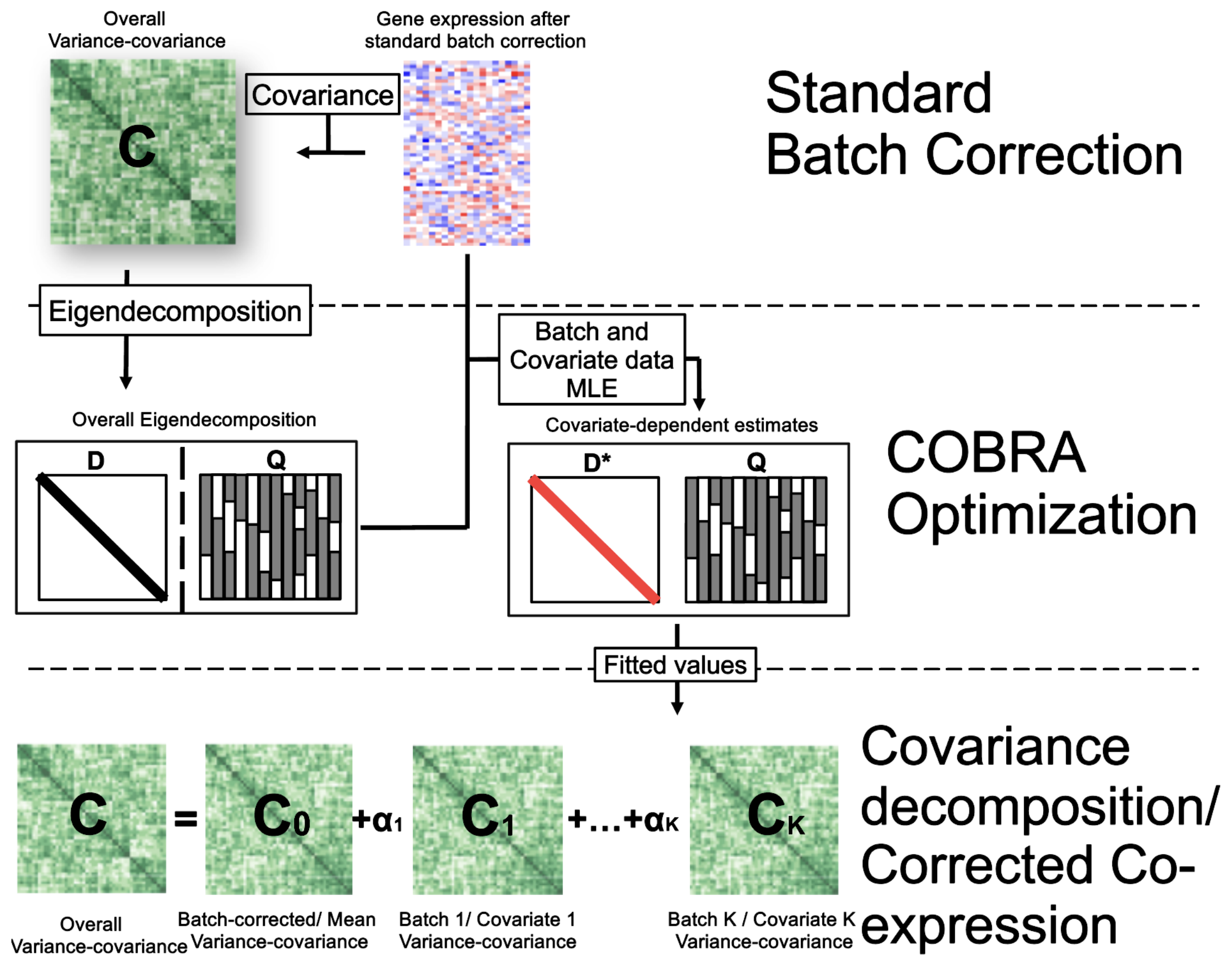
## Supplemental information

← **bioRχiv preprint** (Dec, 2023)
← References & acknowledgements
← Contact information
Tools from our group (`netZoo`) →

## References

1. Johnson et al. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127.
2. Ritchie et al. (2015). Limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47.
3. Leek et al. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 11(10):733–739.
4. Freytag et al. (2015). Systematic noise degrades gene co-expression signals but can be corrected. *BMC bioinformatics*, 16:1–17.
5. Pickrell et al. (2010). Understanding mechanisms underlying human gene expression variation with rna sequencing. *Nature*, 464(7289):768–772.

## The method


Overall Variance-covariance
Gene expression after standard batch correction
C
Covariance
Standard Batch Correction
Eigendecomposition
Batch and Covariate data MLE
Overall Eigendecomposition
D
Q
Covariate-dependent estimates
D*
Q
COBRA Optimization
Fitted values
C = C0 +α1 C1 +...+αK CK
Overall Variance-covariance
Batch-corrected/ Mean Variance-covariance
Batch 1/ Covariate 1 Variance-covariance
Batch K / Covariate K Variance-covariance
Covariance decomposition/ Corrected Co-expression

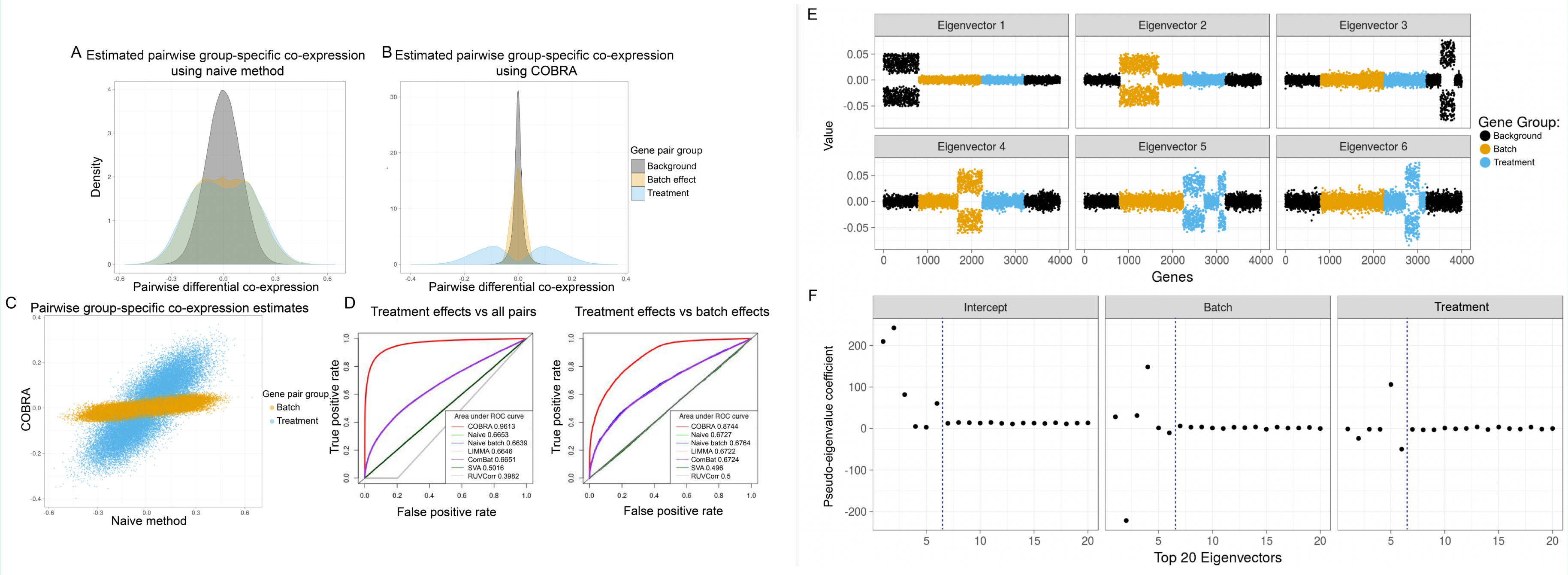
COBRA = **CO**-expression**B**atch**R**eduction**A**djustment integrates a design matrix $\boldsymbol{X}$ to solve

$$\arg\min_{\boldsymbol{\Psi}} ||\boldsymbol{C} - \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{Q} diag(\boldsymbol{X}_i^T \boldsymbol{\Psi})\boldsymbol{Q}^T||_F^2$$

$\hat{\boldsymbol{\Psi}}$ yields a decompostion of $\boldsymbol{C}$ with a component for every covariate in $\boldsymbol{X}$.

| Goal | Design matrix | Output |
|---|---|---|
| Batch corrected case/control comparison | Column 0: intercept<br>Column 1: 0 for control, 1 for case<br>Columns 2...$K$: Batch and covariates | $C_1$ |
| Batch correction<br>&<br>covariate-specific co-expression | Column 0: intercept<br>Columns 1...$K$: covariates of interest | $C_k$ for $k \in [K]$ |

## Improved co-expression estimates *in-silico*

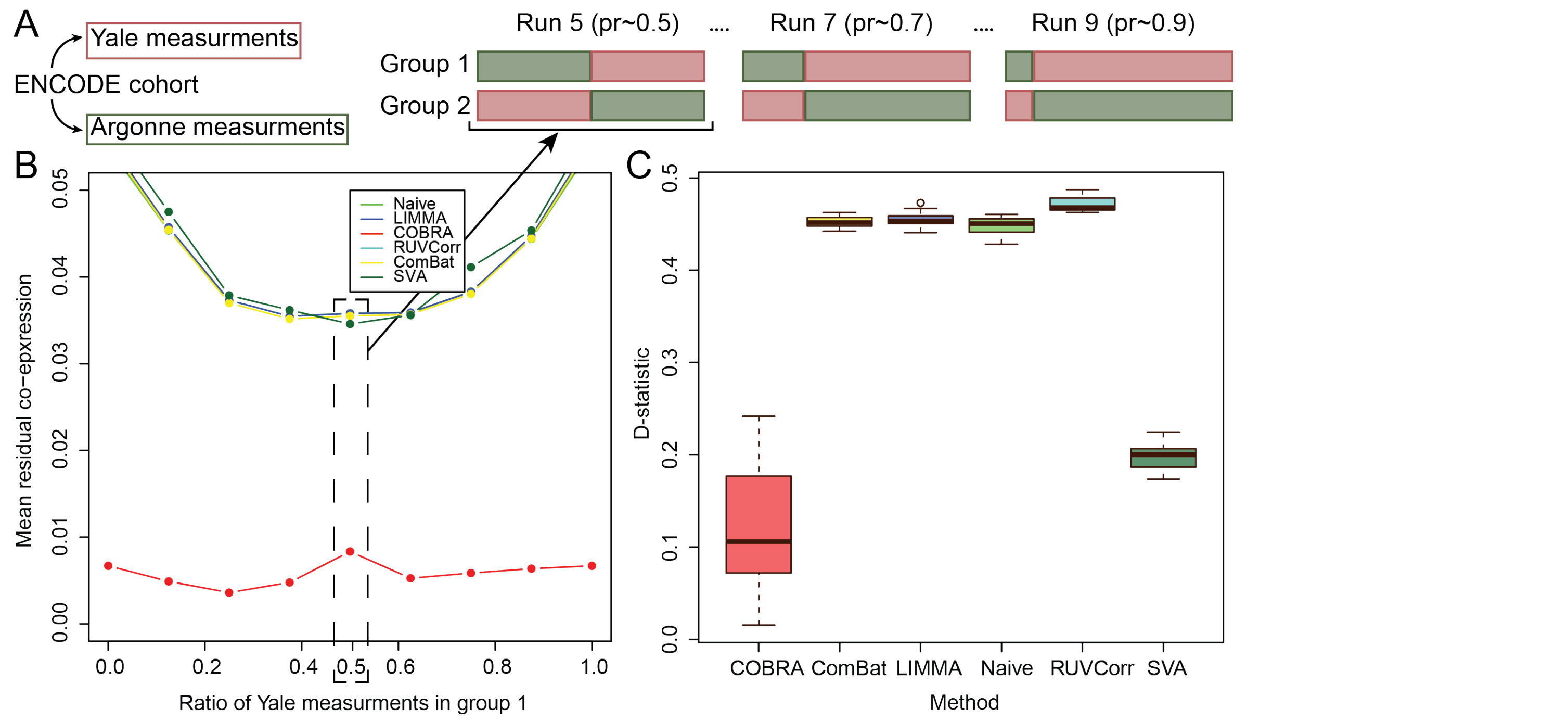


➜ Generated *in-silico* data with known true differential co-expression and batch differential co-expression.

✯ **COBRA** is able to discriminate real effects from batch effects, allowing **effective batch correction for differential co-expression analysis**.

✯ **Interpretation:** $\hat{\boldsymbol{\Psi}}_{i,j}$ is the additional contribution of the $i-$th eigenvector for a one unit increase of the $j-$th covariate.

## Correction of batch effects in ENCODE

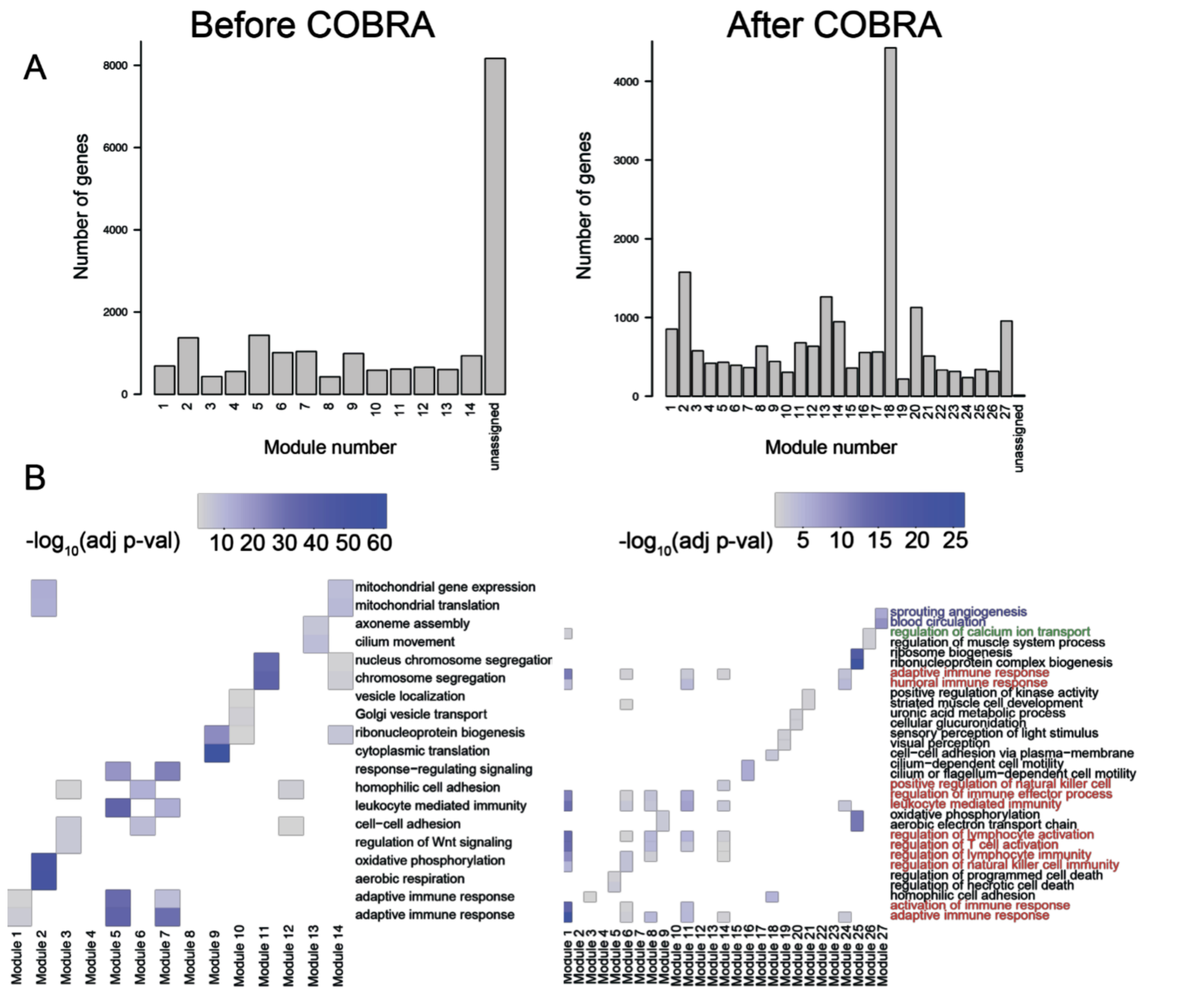


➜ Since the differences between groups are induced by batch, we expect not to see group-specific differential co-expression.

✯ COBRA substantially reduces differential co-expression, **vastly improving** on other methods.

✯ COBRA is more stable across measurement proportions from each lab ⇒ more **robust** estimates.

## Analysis of co-expression modules in thyroid cancer



➜ Applied COBRA to thyroid cancer data from TCGA (controlling for sex, race, stage, batch, and age) and performed GO/ KEGG GSEA on WGCNA modules.

✯ COBRA finds more fine-grained community structures and **facilitates the discovery** of biologically meaningful pathways.

✯ COBRA is not limited to gene co-expression, but it can be **effectively** applied to **partial correlation** networks or as pre-processing of **GRN** inference.