Removing batch-effects from expression data

Maarten van Iterson

Leiden University Medical Center
Department of Molecular Epidemiology

June 23, 2017

Batch-Effects

In differential expression analyses there are primary variables of interest and often other nuisance factors, technical or biological, that introduce unwanted variation.

Batch-Effects

In differential expression analyses there are primary variables of interest and often other nuisance factors, technical or biological, that introduce unwanted variation.

- biological nuisance factors:
  - gender, age, white blood cell composition, etc.

Batch-Effects

In differential expression analyses there are primary variables of interest and often other nuisance factors, technical or biological, that introduce unwanted variation.

- biological nuisance factors:
  - gender, age, white blood cell composition, etc.

- technical nuisance factors (batch effects):
  - lab, sequence machine, library generation date, operator, etc.

Batch-Effects

In differential expression analyses there are primary variables of interest and often other nuisance factors, technical or biological, that introduce unwanted variation.

- biological nuisance factors:
  - gender, age, white blood cell composition, etc.
- technical nuisance factors (batch effects):
  - lab, sequence machine, library generation date, operator, etc.
- Often not all factors are known!

Batch-Effects

In differential expression analyses there are primary variables of interest and often other nuisance factors, technical or biological, that introduce unwanted variation.

- biological nuisance factors:
  - gender, age, white blood cell composition, etc.
- technical nuisance factors (batch effects):
  - lab, sequence machine, library generation date, operator, etc.
- Often not all factors are known!

Confounding occurs when there is correlation between primary variable of interest and the outcome

GEUVADIS RNAseq data[1]

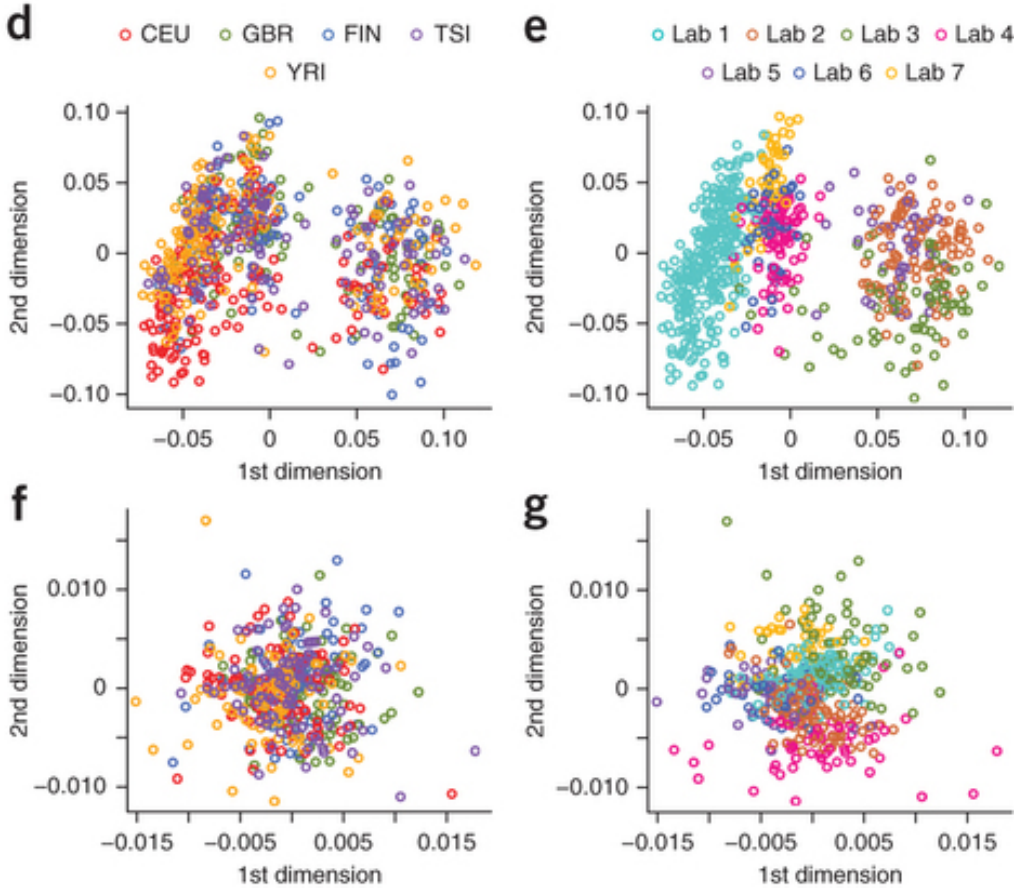

Figure 1 : (d) MDS plot of RNAseq data before batch correction colored by population and (e) colored by laboratory, (f) after batch correction colored by population and (e) colored by laboratory.

---

[1] 't Hoen, P. A. et al. (2013). Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories.
Nat. Biotechnol. 31(11):1015–1022.

Batch correction methods

- normalization methods:
  - quantile normalization, trimmed mean of M-values (TMM)
    edgeR

---

[1] Hansen, K. D., Irizarry, R. A., and Wu, Z. (2012). Removing technical variability in RNA-seq data using conditional quantile normalization.
Biostatistics, 13(2):204–216

[2] Risso, D., Schwartz, K., Sherlock, G., and Dudoit, S. (2011). GC-content normalization for RNA-Seq data.
BMC Bioinformatics, 12:480

Batch correction methods

- normalization methods:
  - quantile normalization, trimmed mean of M-values (TMM) *edgeR*

- technology specific:
  - within-plate-, print-tip-normalization, etc.
  - GC-bias correction methods *cqn*[1], *EDASeq*[2]

---

[1] Hansen, K. D., Irizarry, R. A., and Wu, Z. (2012). Removing technical variability in RNA-seq data using conditional quantile normalization.
Biostatistics, 13(2):204–216

[2] Risso, D., Schwartz, K., Sherlock, G., and Dudoit, S. (2011). GC-content normalization for RNA-Seq data.
BMC Bioinformatics, 12:480

Batch correction methods

- normalization methods:
  - quantile normalization, trimmed mean of M-values (TMM) *edgeR*

- technology specific:
  - within-plate-, print-tip-normalization, etc.
  - GC-bias correction methods *cqn*[1], *EDASeq*[2]

- Batch correction methods:
  - **Nuisance factors are known**: linear model, ComBat
  - **Nuisance factors are unknown**: estimate batch-effects from the data
    - controls e.g. spike-ins or housekeeping
    - principal components
    - ...

---

[1] Hansen, K. D., Irizarry, R. A., and Wu, Z. (2012). Removing technical variability in RNA-seq data using conditional quantile normalization.
Biostatistics, 13(2):204–216

[2] Risso, D., Schwartz, K., Sherlock, G., and Dudoit, S. (2011). GC-content normalization for RNA-Seq data.
BMC Bioinformatics, 12:480

Normalization does not remove batch-effects

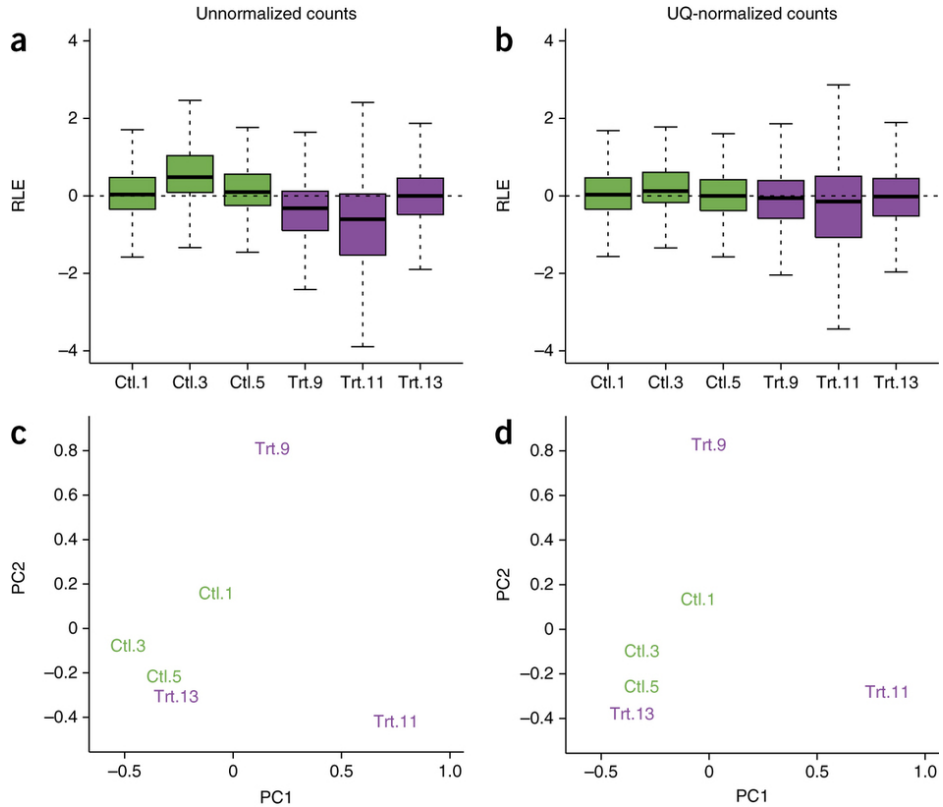

Figure 2 : Raw vs upper-quartile-normalized data [1]

[1] Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples.
Nat. Biotechnol., 32(9):896–902

Removing batch-effects using RUV

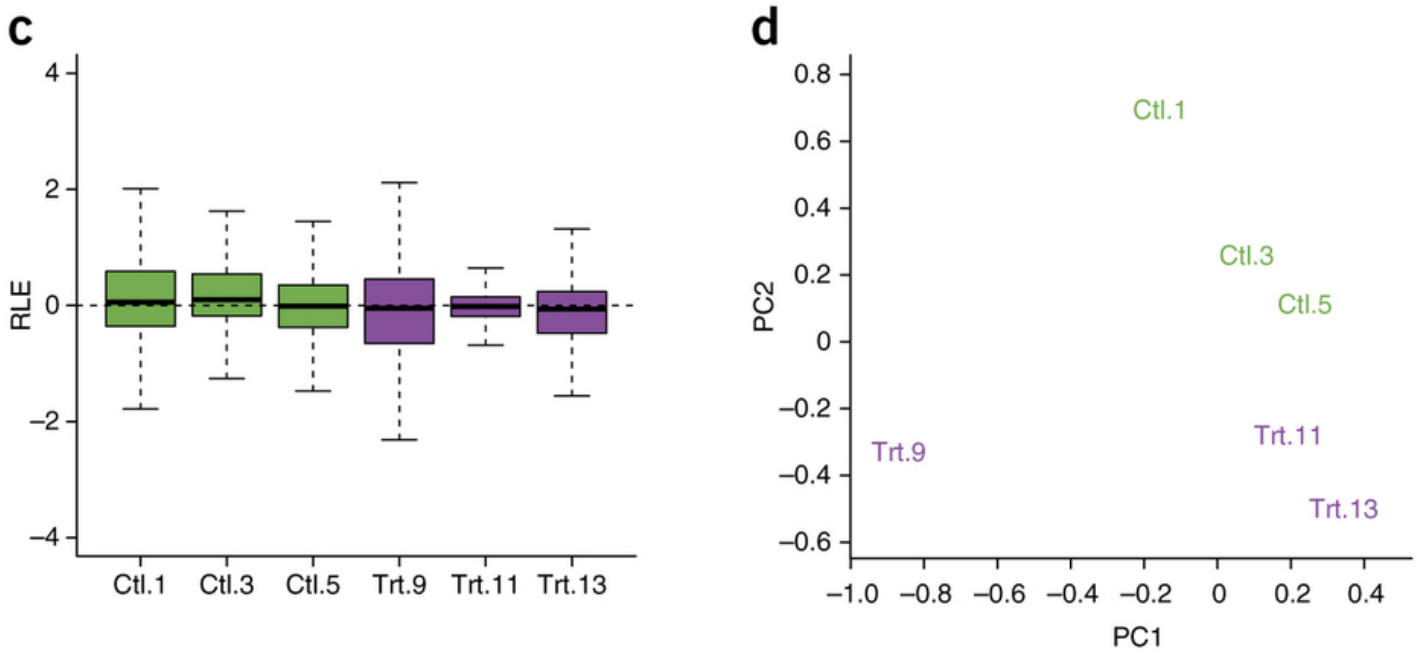

Figure 3 : RUV estimate and corrected

ComBat[1]

Usage:

- Input: Known batches
- Output: Batch corrected expression matrix

[1] Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods.
Biostatistics, 8(1):118–127

ComBat[1]

Usage:

- Input: Known batches
- Output: Batch corrected expression matrix

Method briefly:

- Mean center and standardize the variance of each batch for each gene independently
- Use an empirical Bayes approach to estimate robust mean and variance

[1] Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods.
Biostatistics, 8(1):118–127

ComBat[1]

Usage:

- Input: Known batches
- Output: Batch corrected expression matrix

Method briefly:

- Mean center and standardize the variance of each batch for each gene independently
- Use an empirical Bayes approach to estimate robust mean and variance

Remarks:

- Specially suited for small sample microarray studies
- Method is based on the same idea's for hypothesis testing as implemented in *limma*

R implementation available within the *sva* package

---

[1] Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods.
Biostatistics, 8(1):118–127

Surrogate variable analysis[1]

Usage:

- Input: Does not use known factors but estimates a set of *surrogate variables*
- Optimize: number of surrogate variables
- Output: Estimated surrogate variables
- Testing: Include surrogate variables in a (generalized) linear model

[1] Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis.
PLoS Genetics, 3(9):1724–1735

Surrogate variable analysis[1]

Usage:

- Input: Does not use known factors but estimates a set of *surrogate variables*
- Optimize: number of surrogate variables
- Output: Estimated surrogate variables
- Testing: Include surrogate variables in a (generalized) linear model

Method briefly:

- Constructs *surrogate variables* from a set of genes that are not associated with the biological factor of interest but are affected by unknown batches: principal component analysis on the residuals
- *R* implementation available *sva* package

---

[1] Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis.
PLoS Genetics, 3(9):1724–1735

Removing unwanted variation (RUV)[1]

Usage:

- Input: Does not use known factors but estimates a set of factors describing the *unwanted variation*
- Optimize: Number of unknown factors
- Output: Estimated batch-effects
- Testing: Include estimated batch-effects in a (generalized) linear model

Method briefly:

[1] Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples.
Nat. Biotechnol., 32(9):896–902

Removing unwanted variation (RUV)[1]

Usage:

- Input: Does not use known factors but estimates a set of factors describing the *unwanted variation*
- Optimize: Number of unknown factors
- Output: Estimated batch-effects
- Testing: Include estimated batch-effects in a (generalized) linear model

Method briefly:

- Factor analysis (PC) on the residuals of the control genes
- *R* implementation available *RUVseq*

---

[1] Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples.
Nat. Biotechnol., 32(9):896–902

CATE[1]

Usage:

- Input: Does not use known factors but estimates a set of *latent factors* describing the *unobserved confounding factors*
- Optimize: Number of latent factor
- Output: Estimated latent factors
- Testing: hypotheses testing included (robust regression)

[1] Wang, J., Zhao, Q., Hastie, T., and Owen, A. B. (2015). Confounder Adjustment in Multiple Hypothesis Testing.
ArXiv e-prints

CATE[1]

Usage:

- Input: Does not use known factors but estimates a set of *latent factors* describing the *unobserved confounding factors*
- Optimize: Number of latent factor
- Output: Estimated latent factors
- Testing: hypotheses testing included (robust regression)

Method briefly:

- Factor analysis on *residuals*
- *R* implementation available *cate*

---

[1] Wang, J., Zhao, Q., Hastie, T., and Owen, A. B. (2015). Confounder Adjustment in Multiple Hypothesis Testing.
ArXiv e-prints

Comparison from Leek[1]

Simulated data with one group (Case/Control) and one batch

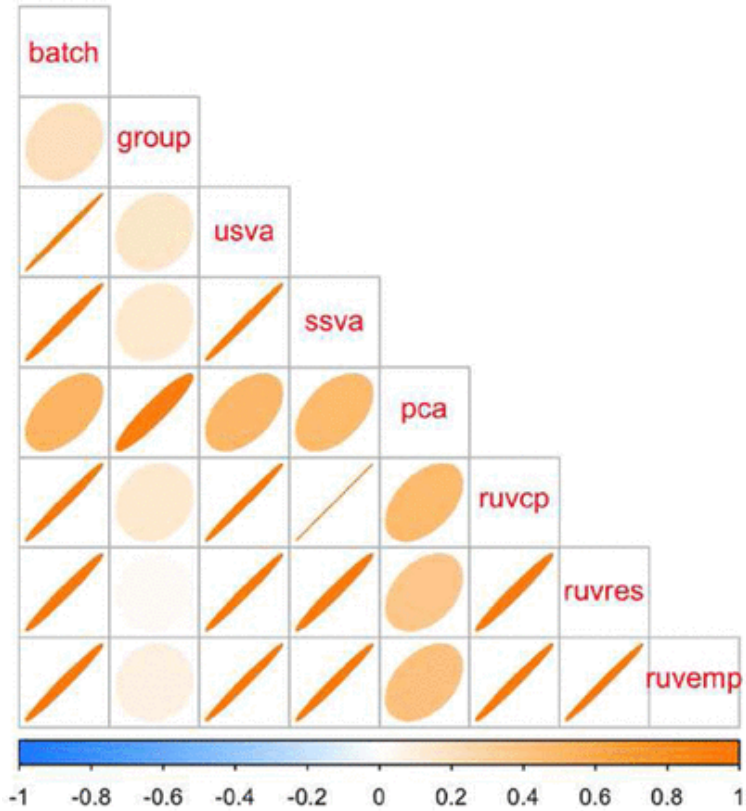

Figure 4 : Correlation between simulated batch and group variables and various batch estimates.

---

[1] Leek, J. T. (2014). svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res.*, 42(21)

Comparison from Leek

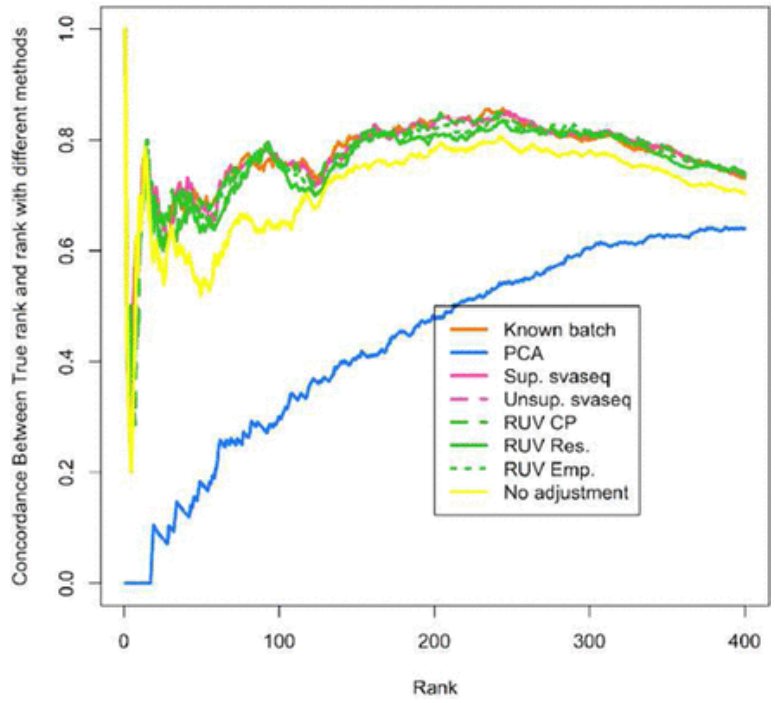

Figure 5 : Differential expression results for simulated data. A concordance at the top plot (CAT plot) shows the fraction of DE results that are concordant between the analysis with the true batch and the analyses using different batch estimates.

A few other methods

1. PEER[1] cran R package *peer*
2. isva[2] cran R package *isva*
3. RUV-4, RUV-inv, and RUV-rinv[3] cran R package *ruv*

These methods can also be applied to other omics-data e.g. 450k DNA methylation data

---

[1] Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc*, 7(3):500–507

[2] Teschendorff, A. E., Zhuang, J., and Widschwendter, M. (2011). Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics*, 27(11):1496–1505

[3] Gagnon-Bartsch, J., Jacob, L., and Speed, T. (2013). Removing unwanted variation from high dimensional data with negative controls. *Tech Report.*

Background: General formulation of the problem

$$Y_{n\times p} = Z_{n\times q}\alpha_{q\times p} + W_{n\times k}\beta_{k\times p} + \epsilon_{n\times p} \quad (1)$$

- $Y_{n\times p}$: observed expr. data on n samples and p genes
- $Z_{n\times q}$: q known cov. including phenotype of interest
- $V_{n\times k}$: k unobserved cov.
- α, β: represent the unknown effects of the obs. and unobs. cov. on gene expr.

Background: Two extremes

Consider only known covariates or no covariates at all

1. Correct for known technical of biological covariates using a linear model:

$$Y = Z\alpha + \epsilon \tag{2}$$

(e.g. all β's are zero)

2. Estimate batch effects using SVD/PC on $Y = U\Sigma V^T$

$$Y = W\beta + \epsilon, \tag{3}$$

$W_{n\times k} = V_{n\times k}$: are the first k principal components, the optimal k needs to be determined (e.g. now all α's are zero)

Background: SVA

Step 1: fit covariates of interest

$$Y_{n\times p} = Z_{n\times 1}\alpha_{q\times p} + \epsilon_{n\times p} \tag{4}$$

$$R_{n\times p} = Y_{n\times p} - Z_{n\times 1}\hat{\alpha}_{q\times p} \tag{5}$$

Estimate unobserved batch effects using SVD/PC on $R = U\Sigma V^T$

Step 2: fit covariates of interest plus a few estimated unobserved batch effects

$$Y_{n\times p} = Z_{n\times q}\alpha_{q\times p} + W_{n\times r}\beta_{r\times p} + \epsilon_{n\times p} \tag{6}$$

$W_{n\times k} = V_{n\times k}$: are the first k principal components, the optimal k is determined using permutation test on the eigenvalues

Background: RUV-2

Step 1: Consider general model for p' control genes

$$Y_{n\times p'} = Z_{n\times 1}\alpha_{q\times p'} + +W_{n\times k}\beta_{k\times p'} + \epsilon_{n\times p'} \tag{7}$$

the 'control gene assumption' is that $\alpha = 0$

$$Y_{n\times p'} = W_{n\times k}\beta_{k\times p'} + \epsilon_{n\times p'} \tag{8}$$

$W_{n\times k} = V_{n\times k}$ SVD/PC of $Y_{n\times p'} = U\Sigma V^T$

Step 2: fit covariates of interest plus a few estimated unobserved batch effects

$$Y_{n\times p} = Z_{n\times q}\alpha_{q\times p} + W_{n\times k}\beta_{k\times p} + \epsilon_{n\times p} \tag{9}$$

the optimal k needs to be determined

if $\alpha \neq 0$ for some covariates use the SVD of $Y - YZ(Z^TZ)^{-1}Z^T$ again only on the control genes

CATE

Background: RUV-4, RUV-inv, and RUV-rinv

RUV-4: a hybrid between RUV-2 and SVA

RUV-inv: includes selection of the number of unobserved batches with a novel method for estimation of the variance $\cdots$

RUV-rinv: rigde regression $\cdots$

Background: isva

i.s.o. using linear uncorrelated eigenvectors in the SVA method estimate statistically independent variables using independent component analysis

Background: PEER

similar to SVA estimates unobserved covariates but uses empirical Bayesian statistics

Relation to other methods: LMM