COMPARISON OF CONFOUND ADJUSTMENT METHODS IN THE CONSTRUCTION OF GENE CO-EXPRESSION NETWORKS

A. Cote, H. Young, L. Hauckins

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

Presentation by Soel Micheletti · August 2023

Genomic data are typically produced in batches,

Genomic data are typically produced in batches, but technical variation and non-biological differences across batches can have unfavorable outcome analyses.

Genomic data are typically produced in batches, but technical variation and non-biological differences across batches can have unfavorable outcome analyses.

The effects of confounding adjustment in gene co-expression analysis are not well understood

KNOWN COVARIATES

PRINCIPAL COMPONENTS (SVA)

RUVCorr

CONFETI

PEER

KNOWN COVARIATES

PRINCIPAL COMPONENTS (SVA)

RUVCorr

CONFETI

PEER

1. For each gene, compute the variance of each covariate on the gene with the *variance-partition* package

2. For covariates accounting for more than 1% of the variance in at least 10% of genes, their effect is regressed out using a linear model

KNOWN COVARIATES

PRINCIPAL COMPONENTS (SVA)

RUVCorr

CONFETI

PEER

Addressing confounding artifacts in reconstruction of gene co-expression networks

Princy Parsana[1†], Claire Ruberman[2†], Andrew E. Jaffe[2,3,4,5,6], Michael C. Schatz[1,7], Alexis Battle[1,8*] and Jeffrey T. Leek[2,6*]

"For scale-free networks, principal components of a gene expression matrix can consistently identify components that reflect artifacts in the data rather than network relationships."

Addressing confounding artifacts in reconstruction of gene co-expression networks

Princy Parsana[1†], Claire Ruberman[2†], Andrew E. Jaffe[2,3,4,5,6], Michael C. Schatz[1,7], Alexis Battle[1,8*] and Jeffrey T. Leek[2,6*]

"For scale-free networks, principal components of a gene expression matrix can consistently identify components that reflect artifacts in the data rather than network relationships."

1. Determine the number of "top" principal components via a permutation approach

Addressing confounding artifacts in reconstruction of gene co-expression networks

Princy Parsana[1†], Claire Ruberman[2†], Andrew E. Jaffe[2,3,4,5,6], Michael C. Schatz[1,7], Alexis Battle[1,8*] and Jeffrey T. Leek[2,6*]

"For scale-free networks, principal components of a gene expression matrix can consistently identify components that reflect artifacts in the data rather than network relationships."

1. Determine the number of "top" principal components via a permutation approach

2. For each gene, a a linear regression **lm(gene ~ top PCs)** is computed and the residuals are kept

KNOWN COVARIATES

PRINCIPAL COMPONENTS (SVA)

RUVCorr

CONFETI

PEER

Systematic noise degrades gene co-expression signals but can be corrected

Saskia Freytag[1,2*], Johann Gagnon-Bartsch[3], Terence P. Speed[1,2,3] and Melanie Bahlo[1,2,4]

$$Y = X\beta + W\alpha + \epsilon$$

Systematic noise degrades gene co-expression signals but can be corrected

Saskia Freytag[1,2*], Johann Gagnon-Bartsch[3], Terence P. Speed[1,2,3] and Melanie Bahlo[1,2,4]

$$Y = X\beta + W\alpha + \epsilon$$

- W can be estimated using factor analysis

Systematic noise degrades gene co-expression signals but can be corrected

Saskia Freytag[1,2*], Johann Gagnon-Bartsch[3], Terence P. Speed[1,2,3] and Melanie Bahlo[1,2,4]

$$Y = X\beta + W\alpha + \epsilon$$

- W can be estimated using factor analysis
- The coefficient of the systematic noise is estimated using Ridge regression and regressed out to obtain the corrected gene expression matrix

Systematic noise degrades gene co-expression signals but can be corrected

Saskia Freytag[1,2*], Johann Gagnon-Bartsch[3], Terence P. Speed[1,2,3] and Melanie Bahlo[1,2,4]

$$Y = X\beta + W\alpha + \epsilon$$

- W can be estimated using factor analysis
- The coefficient of the systematic noise is estimated using Ridge regression and regressed out to obtain the corrected gene expression matrix

Requires a set of negative control genes!!

KNOWN COVARIATES

PRINCIPAL COMPONENTS (SVA)

RUVCorr

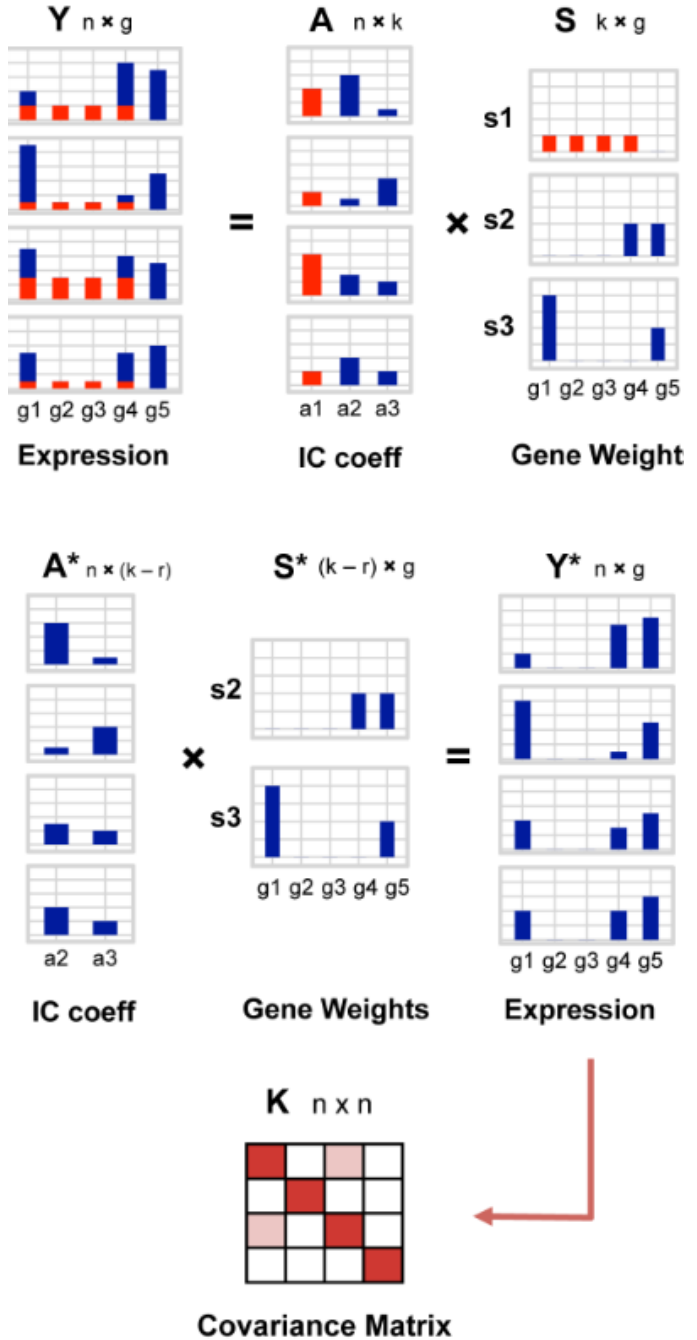
CONFETI

PEER

An independent component analysis confounding factor correction framework for identifying broad impact expression quantitative trait loci

Jin Hyun Ju[1,2], Sushila A. Shenoy[1], Ronald G. Crystal[1], Jason G. Mezey[1,2,3]*

1 Department of Genetic Medicine, Weill Cornell Medical College, New York, NY, United States of America, **2** Institute for Computational Biomedicine, Weill Cornell Medical College, New York, NY, United States of America, **3** Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY, United States of America

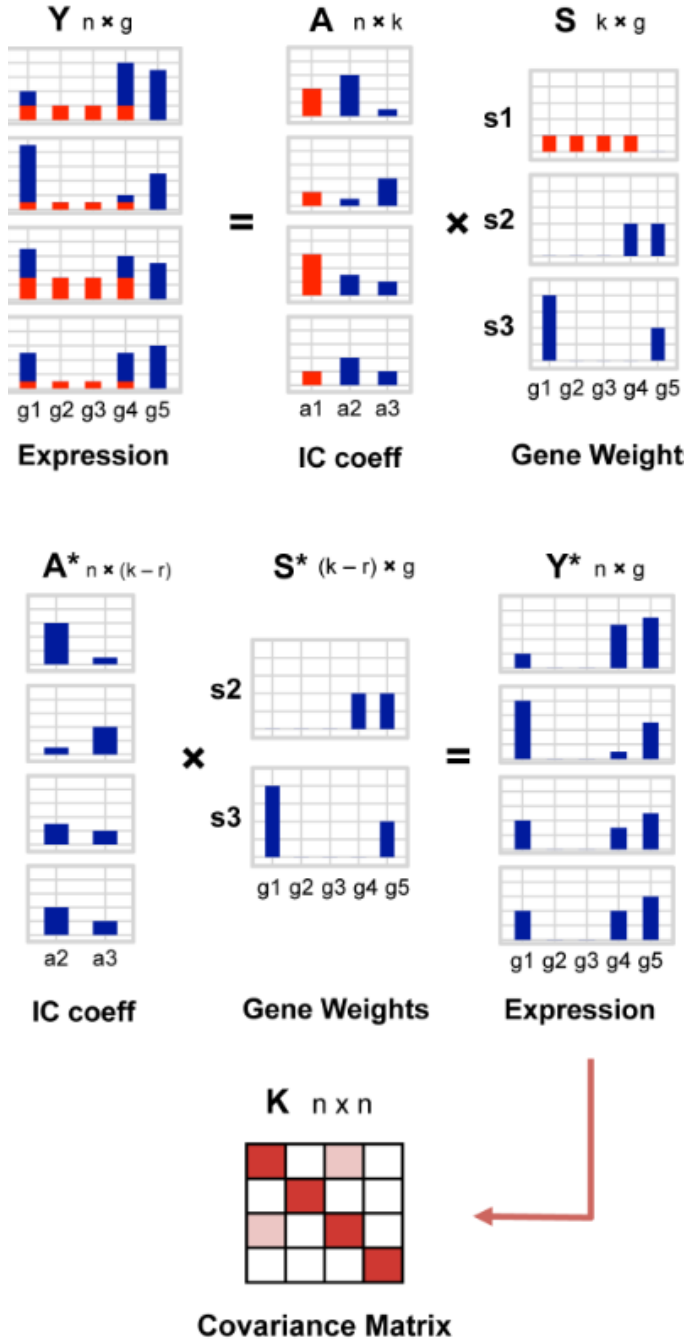


An independent component analysis confounding factor correction framework for identifying broad impact expression quantitative trait loci

Jin Hyun Ju[1,2], Sushila A. Shenoy[1], Ronald G. Crystal[1], Jason G. Mezey[1,2,3]*

1 Department of Genetic Medicine, Weill Cornell Medical College, New York, NY, United States of America, **2** Institute for Computational Biomedicine, Weill Cornell Medical College, New York, NY, United States of America, **3** Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY, United States of America

- ICA is able to more clearly resolve separate factors responsible for variation, while a PCA or factor analysis will tend to identify composite effects.



KNOWN COVARIATES

PRINCIPAL COMPONENTS (SVA)

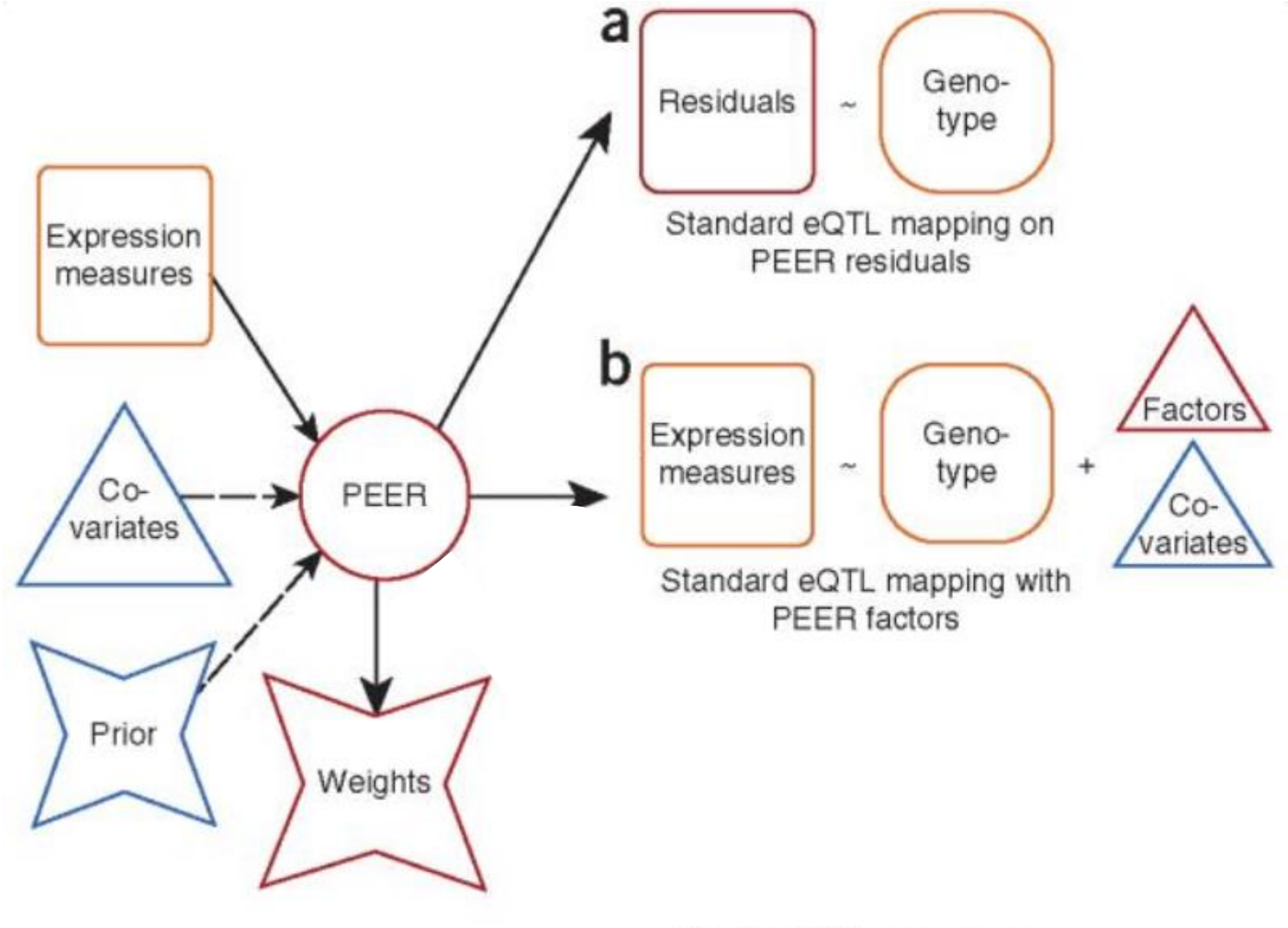
RUVCorr

CONFETI

PEER

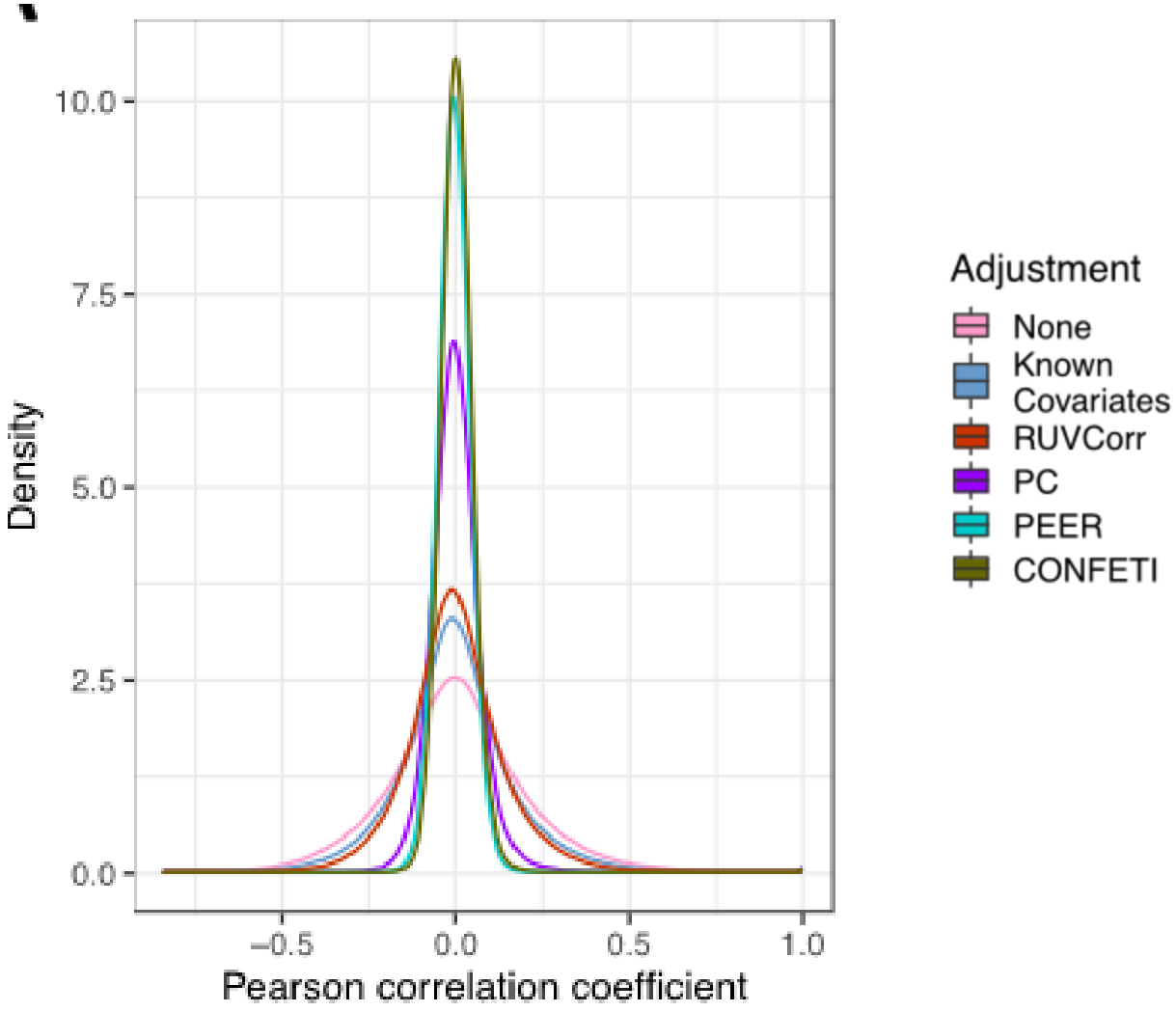
Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses

Oliver Stegle,[1,2,6] Leopold Parts,[3,6] Matias Piipari,[4] John Winn,[5] and Richard Durbin[3]

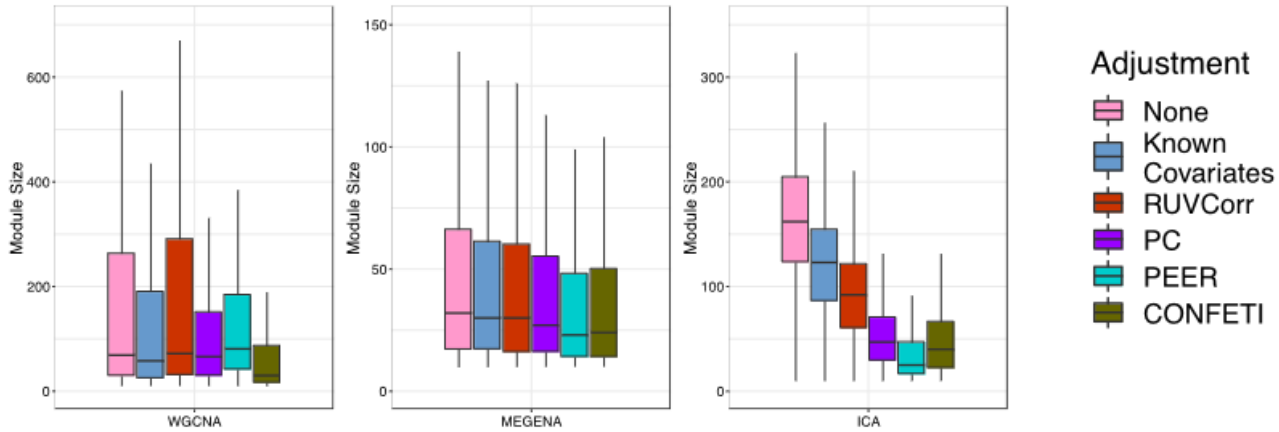


- Computed gene co-expression for 7 GTEx/ CMC tissues
- Hard-thresholding on the absolute co-expression
- Identified modules with WGCNA, MEGENA, and ICA

CONFETI and PEER adjustment result in smaller co-expression networks with fewer gene-gene relationship



Modules identified from CONFETI and PEER-adjusted data tend to be smaller and less variable in size



Overall, there is some overlap among modules identified after known covariate adjustment, RUVCorr, PC, and no data correction (Jaccard index> 0.5) and less overlap between these and modules identified using CONFETI and PEER.

Genome-Scale Integrated Analysis of Networks in Tissues

Genome-Scale Integrated Analysis of Networks in Tissues

- Tissue-specific

Genome-Scale Integrated Analysis of Networks in Tissues

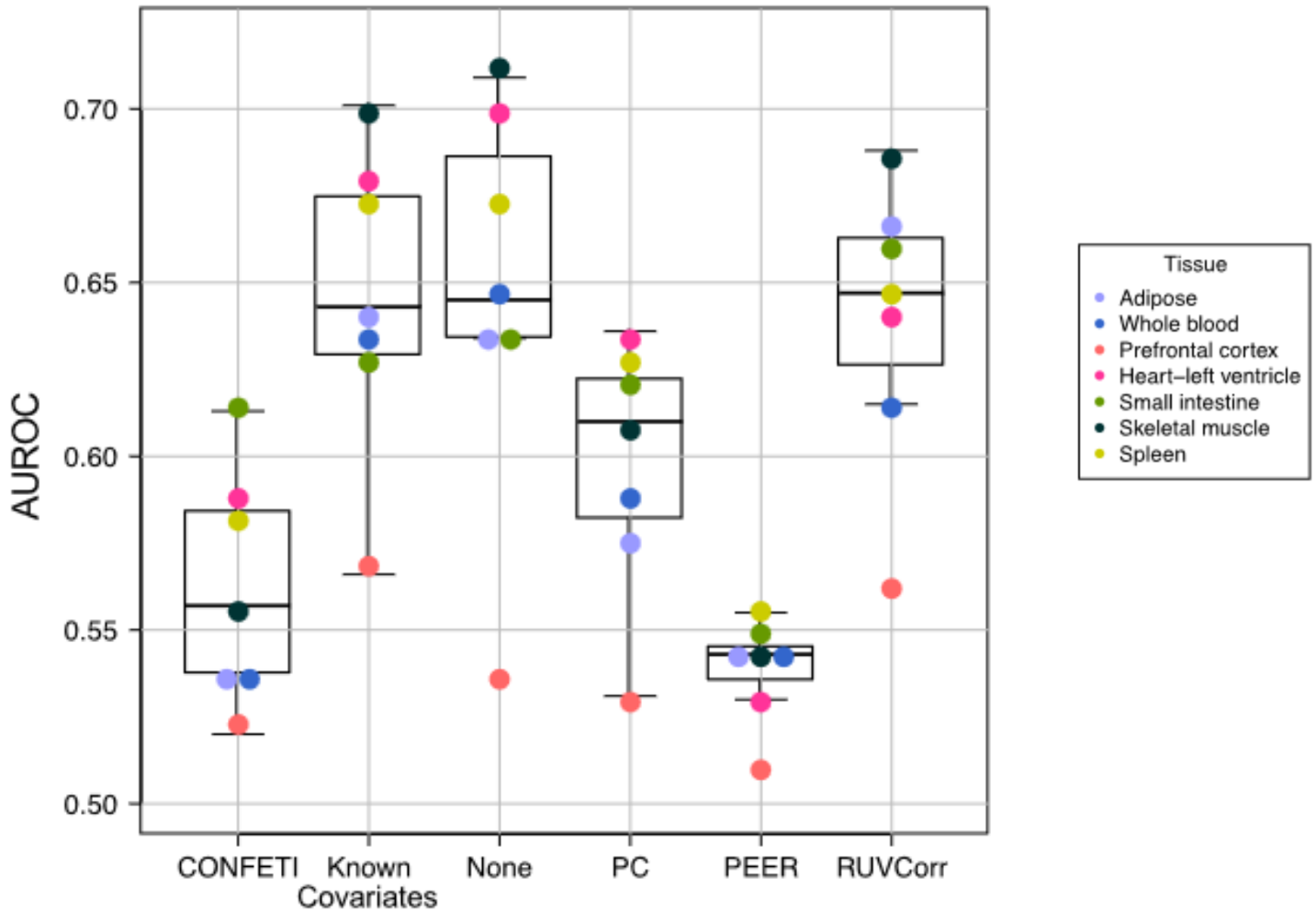
- Tissue-specific
- Pair of genes with functional interactions (high confidence)

Genome-Scale Integrated Analysis of Networks in Tissues

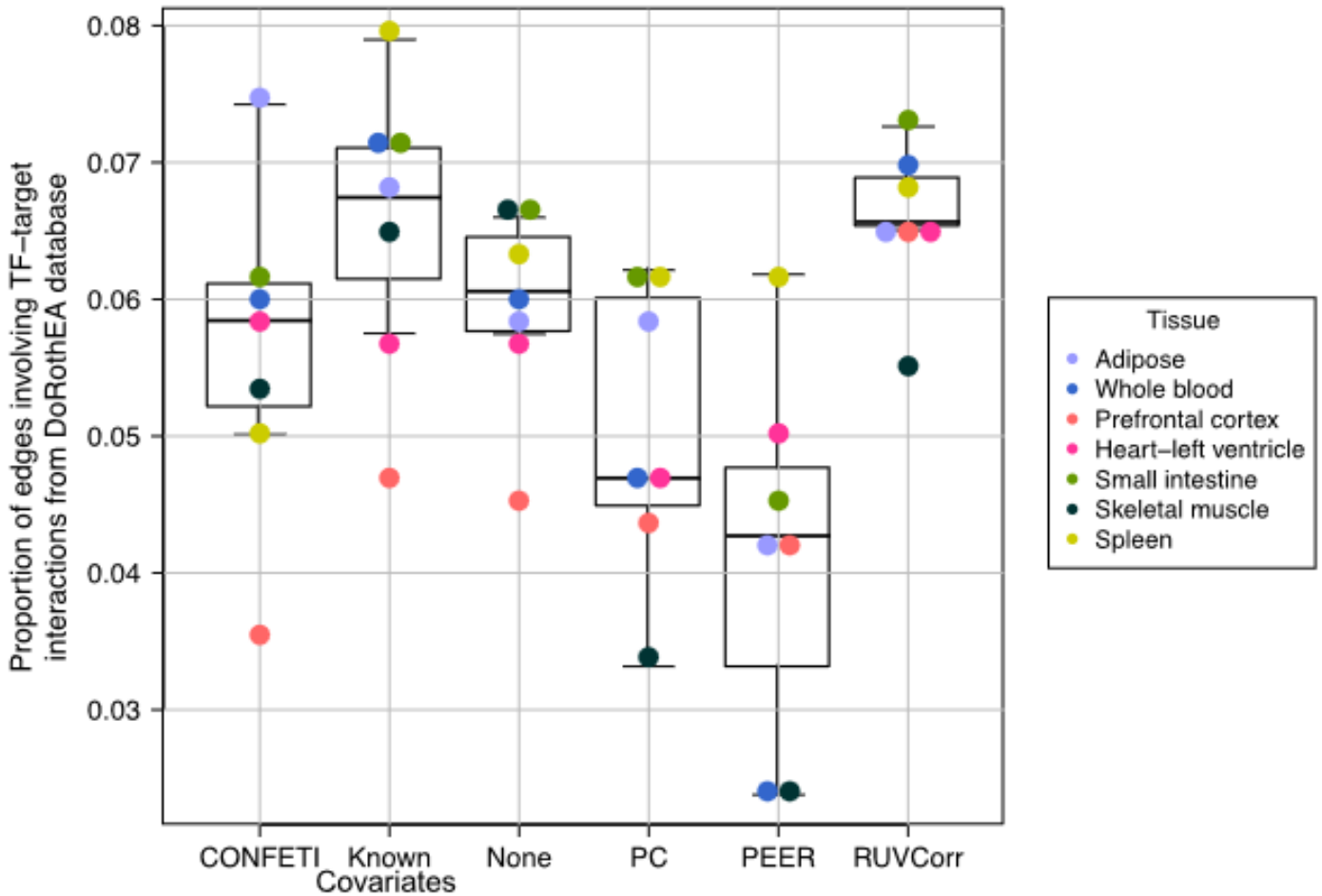
- Tissue-specific
- Pair of genes with functional interactions (high confidence)
- Pair of genes without functional interactions (high confidence)

Genome-Scale Integrated Analysis of Networks in Tissues

- Tissue-specific
- Pair of genes with functional interactions (high confidence)
- Pair of genes without functional interactions (high confidence)
- AUROC on the computed gene co-expression matrices


AUROC
0.70
0.65
0.60
0.55
0.50
CONFETI
Known Covariates
None
PC
PEER
RUVCorr
Tissue
Adipose
Whole blood
Prefrontal cortex
Heart–left ventricle
Small intestine
Skeletal muscle
Spleen

- Tissue agnostic
- TF-target gene relationships from multiple sources
  - ChIP-seq
  - Literature
  - Motif
  - Gene expression


Proportion of edges involving TF–target interactions from DoRothEA database
0.08
0.07
0.06
0.05
0.04
0.03
CONFETI
Known Covariates
None
PC
PEER
RUVCorr
Tissue
Adipose
Whole blood
Prefrontal cortex
Heart–left ventricle
Small intestine
Skeletal muscle
Spleen

- CONFETI and PEER may not be appropriate before co-expression network analysis
  - Very sparse networks
  - Weak representation of known gene-gene interactions

- PC-adjusted datasets show intermediate performance
  - Using many PCs may overcorrect the expression dataset

- RUVCorr correction, known covariate adjustment, and no data correction all performed similarly in this Study

- CONFETI and PEER may not be appropriate before co-expression network analysis
  - Very sparse networks
  - Weak representation of known gene-gene interactions

- PC-adjusted datasets show intermediate performance
  - Using many PCs may overcorrect the expression dataset

- RUVCorr correction, known covariate adjustment, and no data correction all performed similarly in this Study

Although we would theoretically expect that correction at least for known technical factors would improve the accuracy of co-expression networks, there is conflicting evidence that this is the case in practice