香港城市大學
City University of Hong Kong

PCA outperforms popular hidden variable inference methods for molecular QTL mapping

Speaker: LI Yuekai

Date: 07/11/2025

香港城市大學
City University of Hong Kong

PCA outperforms popular hidden variable inference methods for molecular QTL mapping

Speaker: LI Yuekai

Date: 07/11/2025

CONTENTS

香港城市大學
City University of Hong Kong

香港城市大學
City University of Hong Kong

01

Introduction

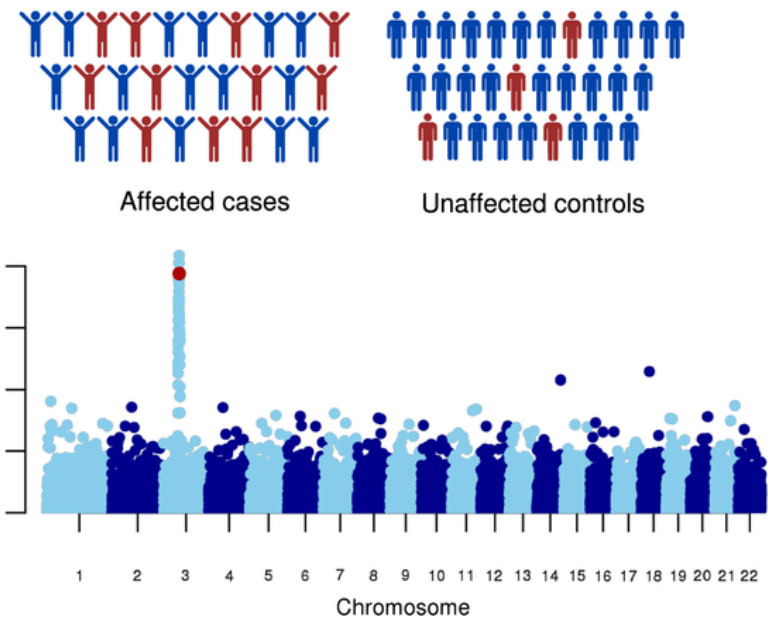
GWAS and complex regulatory mechanisms of SNPs

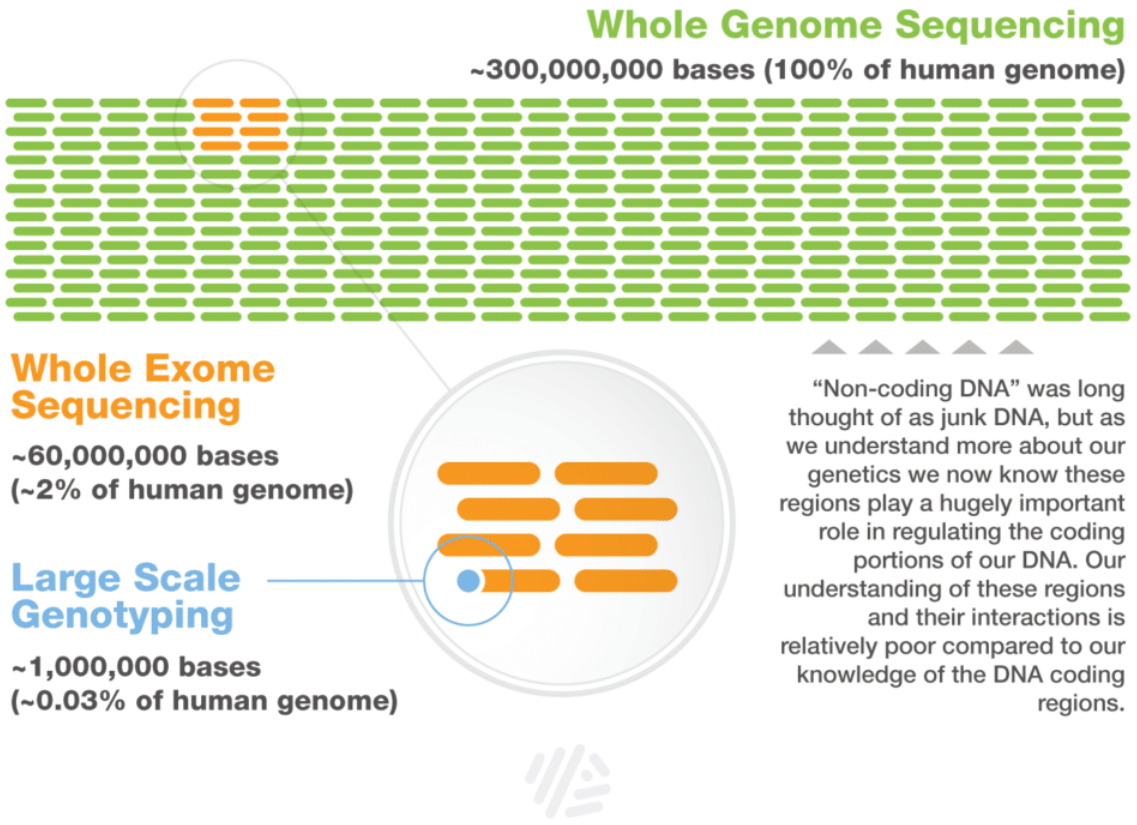
Sequence Variation

ATGCCAGTGTTTCAAGATGCTTGGCCAGCTGGACGAGGGCGATGAC
ATGCCAGTGTTTCAAGATGTTTGGCCAGCTGGACGAGGGCGATGAC

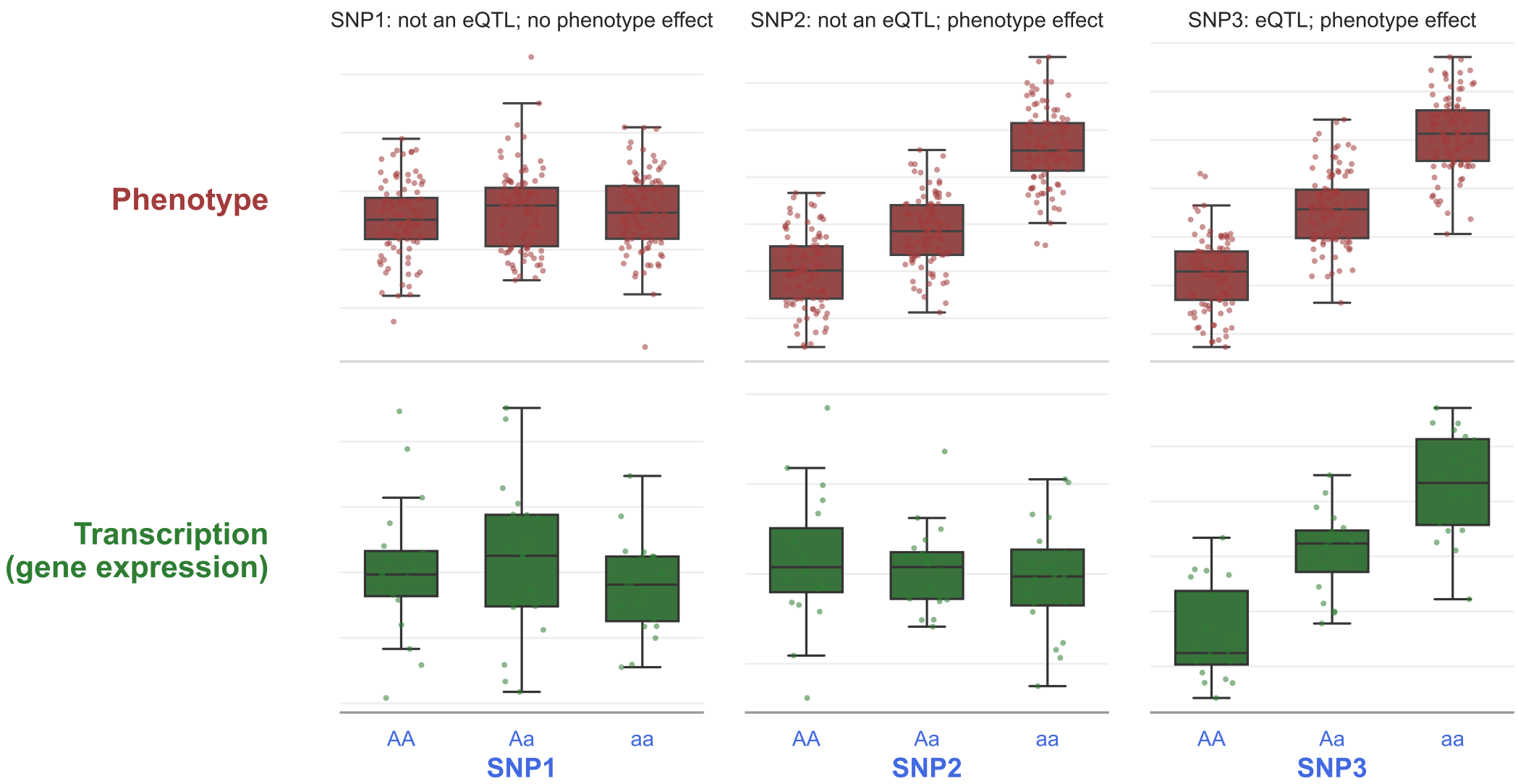
Disease

GWAS

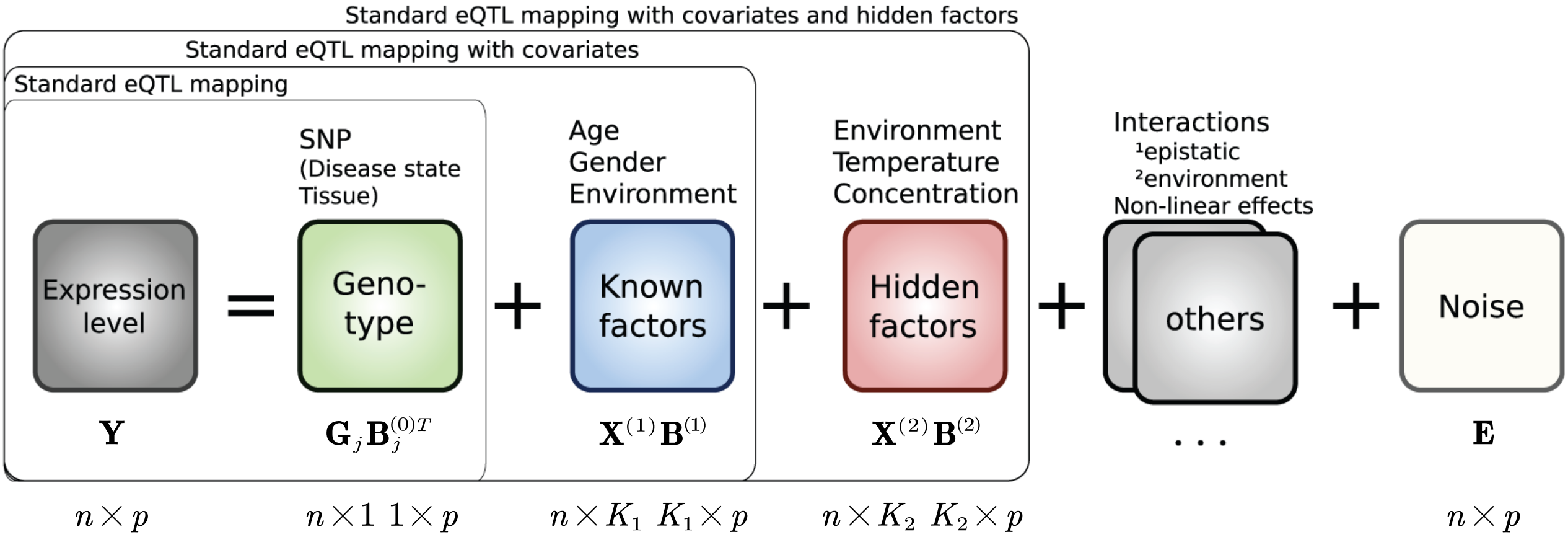




The expression quantitative trait loci (eQTL) analysis



The standard eQTL mapping model



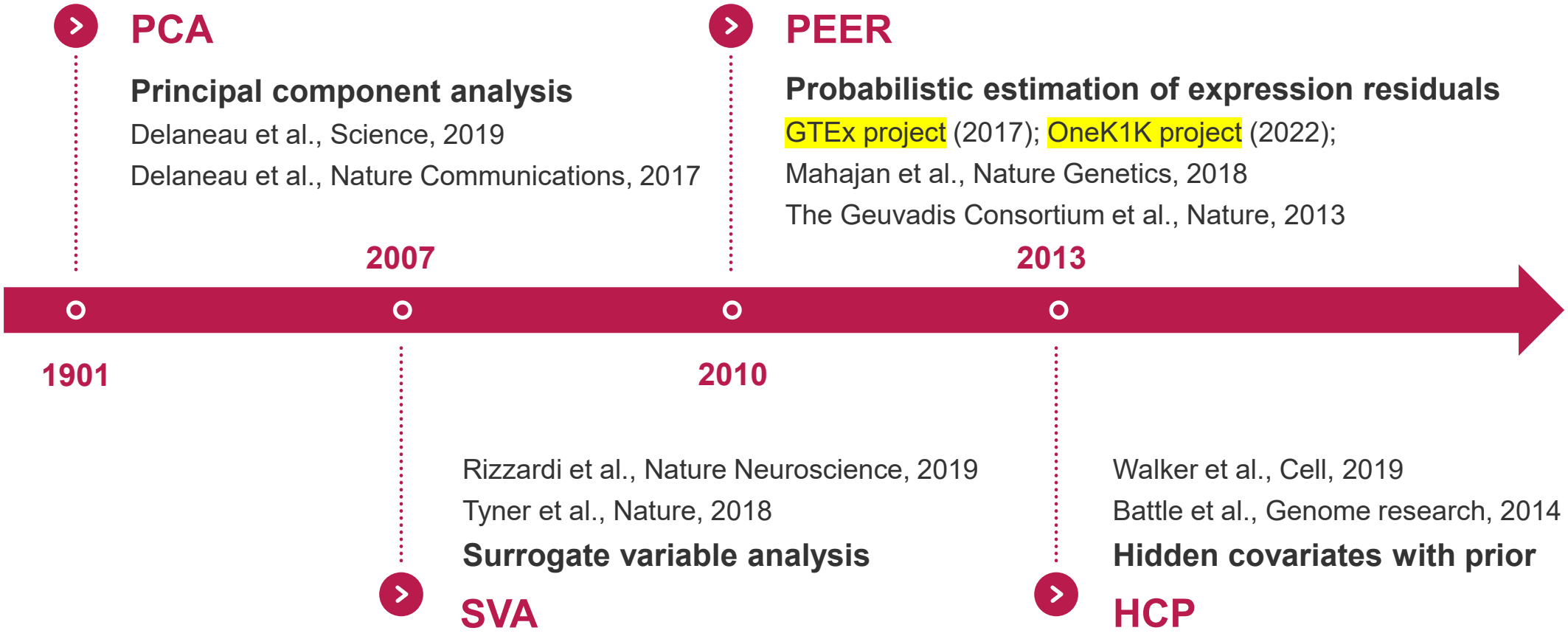
n: sample size

p: number of genes

K_1: number of known covariates

K_2: number of hidden covariates

Some popular hidden variable inference methods for eQTL mapping



香港城市大學
City University of Hong Kong

02

Methods

Principal component analysis (PCA)

Basic idea: Find a loading matrix $\mathbf{W}$, which linearly transforms the observed data $\mathbf{Y}$ into a new feature space.

$$\underset{n\times p}{\mathbf{X}} = \underset{n\times p}{\mathbf{Y}}\ \underset{p\times p}{\mathbf{W}}$$

Constraints: Constrain $\mathbf{W}$ to be orthogonal, and choose its columns $\mathbf{W}_j$ so that each transformed coordinate $\mathbf{YW}_j$ sequentially has the maximum variance.

$$\begin{aligned}
\mathbf{W}_1 &= \underset{\mathbf{W}_1^*\in\mathbb{R}^p}{\operatorname{argmax}} \operatorname{Var}(\mathbf{YW}_1^*) \quad \text{subject to } ||\mathbf{W}_1^*||_2 = 1,\\
\mathbf{W}_2 &= \underset{\mathbf{W}_2^*\in\mathbb{R}^p}{\operatorname{argmax}} \operatorname{Var}(\mathbf{YW}_2^*) \quad \text{subject to } ||\mathbf{W}_2^*||_2 = 1, \mathbf{W}_2^{*T}\mathbf{W}_1 = 0,\\
&\vdots\\
\mathbf{W}_p &= \underset{\mathbf{W}_p^*\in\mathbb{R}^p}{\operatorname{argmax}} \operatorname{Var}(\mathbf{YW}_p^*) \quad \text{subject to } ||\mathbf{W}_p^*||_2 = 1, \mathbf{W}_p^{*T}\mathbf{W}_j = 0, \forall j < p.
\end{aligned}$$

Result: Keep only the first K columns (K < p) of $\mathbf{W}$ to obtain a low-dimensional representation.

$$\underset{n\times K}{\tilde{\mathbf{X}}} = \underset{n\times p}{\mathbf{Y}}\ \underset{p\times K}{\tilde{\mathbf{W}}}$$

Principal component analysis (PCA)

Algorithm 1 Principal component analysis (PCA) algorithm

1: **Input:** Data matrix $\mathbf{Y} \in \mathbb{R}^{n \times p}$ (already standardized); cumulative variance threshold $\tau \in (0, 1)$

2: **Output:** Low–dimension representation $\tilde{\mathbf{X}} \in \mathbb{R}^{n \times K}$; number of components K

3: Compute sample covariance matrix: $\mathbf{S} \leftarrow \dfrac{1}{n-1} \mathbf{Y}^{\mathsf{T}} \mathbf{Y}$

4: Eigendecomposition: $\mathbf{S} = \mathbf{V} \Lambda \mathbf{V}^{\mathsf{T}}$ ▷ $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p),\ \lambda_1 \geq \ldots \geq \lambda_p$

5: Compute cumulative variance ratio: $r_k \leftarrow \sum_{j=1}^{k} \lambda_j / \sum_{j=1}^{p} \lambda_j, \quad k = 1, \ldots, p$

6: Determine smallest K such that $r_K \geq \tau$

7: Select leading eigenvectors: $\tilde{\mathbf{W}} \leftarrow \mathbf{V}_{(:,1:K)}$

8: Project the observed data matrix: $\tilde{\mathbf{X}} \leftarrow \mathbf{Y}\, \tilde{\mathbf{W}}$

9: **return** $\tilde{\mathbf{X}},\ K$

$$\sum_{j=1}^{p} \mathrm{Var}\left(\mathbf{Y}_j\right) = \sum_{j=1}^{p} \lambda_j = \sum_{j=1}^{p} \mathrm{Var}\left(\mathbf{X}_j\right)$$

Factor analysis (FA)

Basic model:

$$\underset{n\times p}{\mathbf{Y}} = \underset{n\times K}{\mathbf{X}}\ \underset{K\times p}{\mathbf{W}} + \underset{n\times p}{\mathbf{E}}$$

Normality assumption:

Hidden factor matrix: $x_{ik} \overset{\text{iid.}}{\sim} \mathcal{N}(0,1), \quad i=1,\ldots,n;\ k=1,\ldots,K$

Error matrix: $e_{ij} \overset{\text{ind.}}{\sim} \mathcal{N}(0,\sigma_j^2),\ i=1,\ldots,n;\ j=1,\ldots,p$

$$\Rightarrow \text{Cov}(x_{ik},\ e_{ij}) = 0, \text{for } \forall i,j,k$$

where:

Y is the observed data, and the marginal distribution of **Y** is normal.

W and σ_j^2 are fixed but unknown parameters

X is the missing data (latent variable)

As shown by Rubin and Thayer (1997), the EM algorithm can be used to solve this maximization problem by treating **X** as missing data.

Factor analysis (FA)

Joint distribution:

Define $\boldsymbol{\psi} = [\sigma_1^2, \ldots, \sigma_p^2]$, we have:

$$\begin{bmatrix} \mathbf{x}_i \\ \mathbf{y}_i \end{bmatrix} \overset{\text{iid.}}{\sim} \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{I}_K & \mathbf{W} \\ \mathbf{W}^T & \mathbf{W}^T\mathbf{W} + \boldsymbol{\Psi} \end{bmatrix}\right), \quad i = 1, \cdots, n$$

EM algorithm:

E-step:

$$\mathbb{E}[\mathbf{x}_i \mid \mathbf{y}_i] = \mathbf{W}\left(\mathbf{W}^T\mathbf{W} + \boldsymbol{\Psi}\right)^{-1}\mathbf{y}_i$$
$$\mathrm{Cov}[\mathbf{x}_i \mid \mathbf{y}_i] = \mathbf{I}_K - \mathbf{W}\left(\mathbf{W}^T\mathbf{W} + \boldsymbol{\Psi}\right)^{-1}\mathbf{W}^T$$
$$\mathbb{E}[\mathbf{x}_i\mathbf{x}_i^T \mid \mathbf{y}_i] = \mathrm{Cov}[\mathbf{x}_i \mid \mathbf{y}_i] + \mathbb{E}[\mathbf{x}_i \mid \mathbf{y}_i]\mathbb{E}[\mathbf{x}_i \mid \mathbf{y}_i]^T$$

→ Thomson's factor scores

M-step:

$$\mathbf{W}^{\text{new}} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\mathbf{x}_i|\mathbf{y}_i]\mathbf{y}_i^T\right)\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\mathbf{x}_i\mathbf{x}_i^T|\mathbf{y}_i]\right)^{-1}$$

$$\boldsymbol{\Psi}^{\text{new}} = \mathrm{diag}\left\{\frac{1}{n}\sum_{i=1}^{n}[\mathbf{y}_i\mathbf{y}_i^T - \mathbf{W}^{\text{new}T}\mathbb{E}[\mathbf{x}_i|\mathbf{y}_i]\mathbf{y}_i^T]\right\}$$

plug in

parameter estimates at convergence

Probabilistic principal component analysis (PPCA)

Basic model:

$$\underset{n\times p}{\mathbf{Y}} = \underset{n\times K}{\mathbf{X}}\ \underset{K\times p}{\mathbf{W}} + \underset{n\times p}{\mathbf{E}}$$

Normality assumption:

$$\left.\begin{aligned} &\text{Hidden factor matrix:} && x_{ik} \overset{\text{iid.}}{\sim} \mathcal{N}(0,1), \quad i=1,\ldots,n;\ k=1,\ldots,K \\ &\text{Error matrix:} && e_{ij} \overset{\text{iid.}}{\sim} \mathcal{N}(0,\sigma^2),\ i=1,\ldots,n;\ j=1,\ldots,p \end{aligned}\right\} \quad \mathrm{Cov}(x_{ik},\ e_{ij}) = 0, \text{for } \forall i,j,k$$

In contrast to factor analysis, the maximum likelihood can be obtained analytically in PPCA:

$$\left.\begin{aligned} \sigma^2_{\mathrm{MLE}} &= \frac{1}{p-K}\sum_{j=K+1}^{p}\lambda_j \\ \mathbf{W}_{\mathrm{MLE}} &= \mathrm{V}_K\left(\Lambda_K - \sigma^2_{\mathrm{MLE}}\mathbf{I}_K\right)^{\frac{1}{2}} \end{aligned}\right\} \Rightarrow \mathbb{E}\left[\mathbf{x}_i|\mathbf{y}_i\right] = \mathbf{W}_{\mathrm{MLE}}\left(\mathbf{W}^T_{\mathrm{MLE}}\mathbf{W}_{\mathrm{MLE}} + \sigma^2_{\mathrm{MLE}}\mathbf{I}_p\right)^{-1}\mathbf{y}_i$$

where, V_k, Λ_K, λ_j are all derived from the eigendecomposition of the sample covariance matrix.

Probabilistic estimation of expression residuals (PEER)

Basic model:

$$\underset{n\times p}{\mathbf{Y}} = \underset{n\times(K_1+K_2)}{[\mathbf{X}^{(1)},\ \mathbf{X}^{(2)}]}\ \underset{(K_1+K_2)\times p}{\mathbf{W}} + \underset{n\times p}{\mathbf{E}}$$

Normality assumption:

Hidden factor matrix: $x_{ik}^{(2)} \overset{\text{iid.}}{\sim} \mathcal{N}(0,1), \quad i=1,\ldots,n;\ k=1,\ldots,K_2$

Covariate-specific weight matrix: $w_{kj} \overset{\text{ind.}}{\sim} \mathcal{N}\left(0, \frac{1}{\beta_k}\right),\ j=1,\ldots,p;\ k=1,\ldots,K_1+K_2$

Error matrix: $e_{ij} \overset{\text{ind.}}{\sim} \mathcal{N}\left(0, \frac{1}{\tau_j}\right),\ i=1,\ldots,n;\ j=1,\ldots,p$

Prior distribution:

$$\beta_k \overset{\text{iid.}}{\sim} \Gamma(a_1,b_1),\ k=1,\ldots,K_1+K_2$$

$$\tau_j \overset{\text{iid.}}{\sim} \Gamma(a_2,b_2),\ j=1,\ldots,p$$

Hidden covariates with prior (HCP)

Basic model:

$$\underset{n\times p}{\mathbf{Y}} = \underset{n\times K_1}{\mathbf{X}^{(1)}}\underset{K_1\times p}{\mathbf{W}^{(1)}} + \underset{n\times K_2}{\mathbf{X}^{(2)}}\underset{K_2\times p}{\mathbf{W}^{(2)}} + \underset{n\times p}{\mathbf{E}}$$

- $\mathbf{X}^{(1)}$ is known covariates;
- $\mathbf{X}^{(2)}$ is unknown covariates;
- $\mathbf{W}^{(1)}$ is fixed effects for known covariates.

Normality assumption:

$$x^{(2)}_{ik_2} \overset{\text{iid.}}{\sim} \mathcal{N}\left((\mathbf{X}^{(1)}\mathbf{C})_{ik_2}, \frac{1}{\lambda_1}\right),\ i=1,\ldots,n;\ k_2=1,\ldots,K_2$$

$$c_{k_1k_2} \overset{\text{iid.}}{\sim} \mathcal{N}\left(0, \frac{1}{\lambda_2}\right),\ k_1=1,\ldots,K_1;\ k_2=1,\ldots,K_2$$

$$w^{(2)}_{k_2j} \overset{\text{iid.}}{\sim} \mathcal{N}\left(0, \frac{1}{\lambda_3}\right),\ j=1,\ldots,p;\ k_2=1,\ldots,K_2$$

$$e_{ij} \overset{\text{iid.}}{\sim} \mathcal{N}\left(0, \frac{1}{\lambda}\right),\ i=1,\ldots,n;\ j=1,\ldots,p$$

The prior of unknown covariates depends on linear combinations of known covariates.

Penalized regression form:

$$\underset{\mathbf{X}^{(2)},\mathbf{C},\mathbf{W}^{(1)},\mathbf{W}^{(2)}}{\operatorname{argmin}} \ ||\mathbf{Y}-\mathbf{X}^{(1)}\mathbf{W}^{(1)}-\mathbf{X}^{(2)}\mathbf{W}^{(2)}||_2^2 + \lambda_1||\mathbf{X}^{(2)}-\mathbf{X}^{(1)}\mathbf{C}||_2^2 + \lambda_2||\mathbf{C}||_2^2 + \lambda_3||\mathbf{W}^{(2)}||_2^2$$

Surrogate variable analysis (SVA)

Algorithm 2-1 Iteratively reweighted surrogate variable analysis (IRW-SVA)

Input:

- Gene expression matrix $\mathbf{Y}$; variables of interest $\mathbf{G}$; known covariates $\mathbf{X}_1$ (optional).
- Number of inferred variables K; number of iterations B.

Output: $\mathbf{X}_2$ ($n \times K$ matrix of inferred covariates).
for each gene **do**
    Regress the j-th column of $\mathbf{Y}$ against $(\mathbf{1}, \mathbf{G}, \mathbf{X}_1)$;
    Replace the j-th column of $\mathbf{Y}$ with the residuals to get $\mathbf{R}$;
end for
Perform PCA on $\mathbf{R}$ to get the initial PCs;
for $b \leftarrow 1$ **to** B **do**
    for each gene **do**
        Test H_0: coefficients of $\mathbf{G} = 0$ in the regression against $(1, \mathbf{G}, \mathbf{X}_1, \text{PCs})$ via partial F-test;
        Convert the p-values to local false discovery rates, denoted as lfdr_1;
    end for
    for each gene **do**
        Test H_0: coefficients of PCs $= 0$ in the regression against $(1, \mathbf{X}_1, \text{PCs})$ via partial F-test;
        Convert the p-values to local false discovery rates, denoted as lfdr_2;
    end for
    Weight the columns of $\mathbf{Y}$ by $\text{lfdr}_1 \times (1 - \text{lfdr}_2)$;
    Perform PCA on the weighted $\mathbf{Y}$ after centering;
end for
return the first K PCs from the last PCA.

SVA is purely algorithmic, based on the PCA algorithm.

Surrogate variable analysis (SVA)

Algorithm 2-1 Iteratively reweighted surrogate variable analysis (IRW-SVA)

Input:

- Gene expression matrix $\mathbf{Y}$; variables of interest $\mathbf{G}$; known covariates $\mathbf{X}_1$ (optional).
- Number of inferred variables K; number of iterations B.

Output: $\mathbf{X}_2$ ($n \times K$ matrix of inferred covariates).
for each gene **do**
  Regress the j-th column of $\mathbf{Y}$ against $(\mathbf{1}, \mathbf{G}, \mathbf{X}_1)$;
  Replace the j-th column of $\mathbf{Y}$ with the residuals to get $\mathbf{R}$;
end for
Perform PCA on $\mathbf{R}$ to get the initial PCs;
for $b \leftarrow 1$ **to** B **do**
  for each gene **do**
    Test H_0: coefficients of $\mathbf{G} = 0$ in the regression against $(1, \mathbf{G}, \mathbf{X}_1, \text{PCs})$ via partial F-test;
    Convert the p-values to local false discovery rates, denoted as lfdr_1;
  end for
  for each gene **do**
    Test H_0: coefficients of PCs $= 0$ in the regression against $(1, \mathbf{X}_1, \text{PCs})$ via partial F-test;
    Convert the p-values to local false discovery rates, denoted as lfdr_2;
  end for
  Weight the columns of $\mathbf{Y}$ by $\text{lfdr}_1 \times (1 - \text{lfdr}_2)$;
  Perform PCA on the weighted $\mathbf{Y}$ after centering;
end for
return the first K PCs from the last PCA.

SVA is purely algorithmic, based on the PCA algorithm.

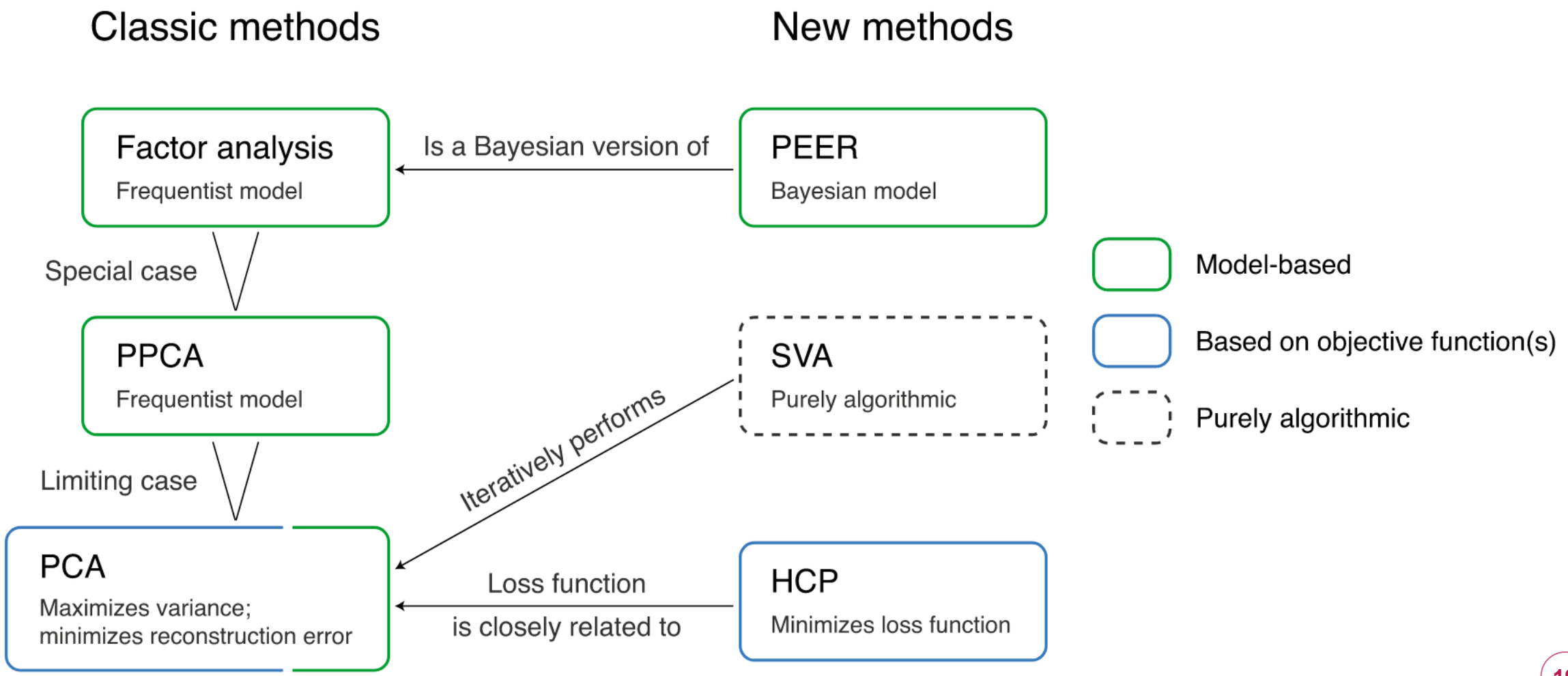
Algorithm 2-2 The calculation for lfdr in IRW-SVA algorithm

Input: p-values pVals; vector of length p; threshold $\lambda = 0.8, \epsilon = 10^{-8}$.
Output: local false discovery rates lfdrs.
$\widehat{\pi}_0 \leftarrow \text{sum}(pVals > \lambda)/p(1-\lambda)$;
Floor each element of pVals at ϵ and cap each element of pVals at $1 - \epsilon$;
$\widehat{f}_0 \leftarrow \text{dnorm}(\text{qnorm}(\text{pVals}))$;
$\widehat{f}$ is obtained by fitting a smooth curve to the histogram of pVals via kernel density estimation;
$\text{lfdrs} \leftarrow \widehat{\pi}_0 \widehat{f}_0 / \widehat{f}$;
return lfdrs;

Algorithm 2-3 the BE algorithm for choosing K in SVA

Input: Gene expression matrix $\mathbf{Y}$; variables of interest $\mathbf{G}$; known covariates $\mathbf{X}_1$ (optional).
Output: Number of hidden covariates K.
Residualize $\mathbf{Y}$ against $(\mathbf{1}, \mathbf{G}, \mathbf{X}_1)$ to obtain $\mathbf{R}$;
Perform PCA on $\mathbf{R}$;
Denote the proportion of variance explained (PVE) by the k-th PC by PVE_k;
for $b \leftarrow 1$ **to** $B = 20$ **do**
  Permute each column of $\mathbf{R}$ to obtain $\mathbf{R}_b$;
  Residualize $\mathbf{R}_b$ against $(\mathbf{1}, \mathbf{G}, \mathbf{X}_1)$ to obtain $\mathbf{R}'_b$;
  Perform PCA on $\mathbf{R}'_b$;
end for
The p-value for the k-th PC is calculated as the proportion of permutations where the PVE of the kth PC is greater than or equal to PVE_k;
Enforce that the p-values increase (i.e., are non-decreasing) as k increases;
return the number of PCs with a p-value smaller than or equal to $\alpha = 0.1$.

The relationship between PCA and hidden variable inference methods



Comparison of PCA, PEER, HCP, and SVA

Method	Variables of interest	Known covariates	K	tuning parameters	probabilistic model
PCA	Unrequired	Unrequired	Adaptive	Unrequired	No
PEER	Unrequired	Optional	Required	Optional	Yes
HCP	Unrequired	Required	Required	Required	Yes
SVA	Required	Optional	Optional	Unrequired	No

香港城市大學
City University of Hong Kong

03

Results

Limitations of the simulation in the original PEER publication

Data analysis limitations:

(a) The study only compares PEER against the other methods in terms of power, not in terms of FPR or FDR.

(b) The study does not use PCA or SVA properly.

(c) The study does not evaluate the different ways of using PEER.

(d) The study uses non-standard priors for PEER that are different from the default priors.

Data simulation limitations:

(a) The data dimensions are minimal, with only q = 100 SNPs in the whole genome.

(b) The SNP genotypes are simulated independently and identically with a target minor allele frequency (MAF) of 0.4, so there is no linkage disequilibrium (LD) and a higher average MAF than in real data.

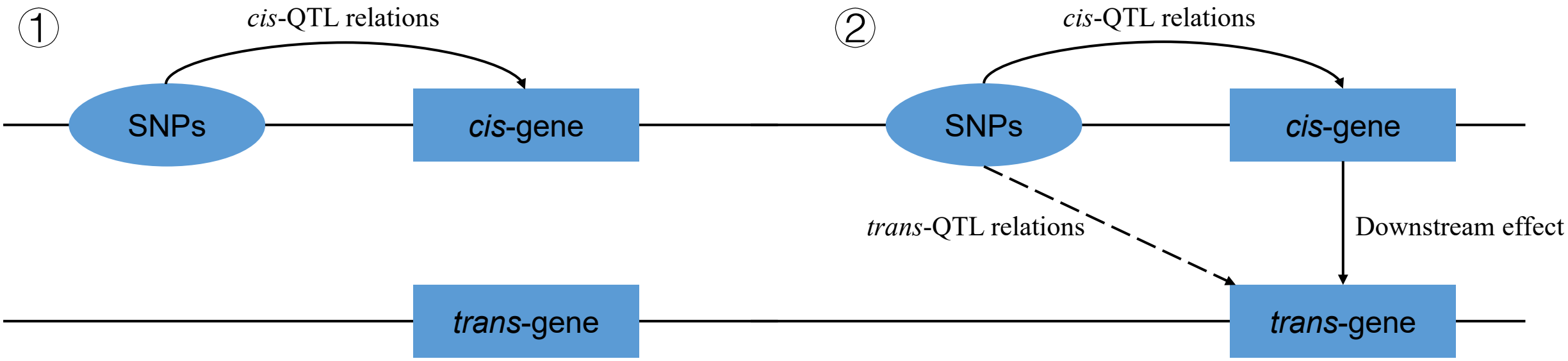
(c) The gene expression levels are primarily driven by *trans*-regulatory effects rather than *cis*-regulatory effects or covariate effects.

(d) The study only simulates one replicate of one experiment.

Simulation design 1 versus Simulation design 2

	Simulation Design 1	Simulation Design 2
Number of experiments	1	176
Number of replicates per experiment	10	2
Genotype data	Simulated (no LD, high MAF)	Real genotype from GTEx
Cis-QTL relations present	✓	✓
Trans-QTL relations present	✓	✗
Number of individuals	n=80	n=838
Number of genes	p=400	p=1000
Number of SNPs	q=100 SNPs in the whole genome	q=1000 local common SNPs per gene
Number of causal SNPs per gene	random	1 or random
Number of known covariates	k_1=3	k_1=2,3,5, or 8, respectively
Number of hidden covariates	k_2=7	k_2=3,7,15, or 22, respectively

Data simulation for trans-QTL detection (design 1)



$$\underset{n\times p}{\mathbf{Y}} = \underset{n\times p}{\mathbf{Y}_{\text{BeforeDSE}}} + \underset{n\times p}{\mathbf{Y}_{\text{DSE}}}$$

$$\underset{n\times p}{\mathbf{Y}_{\text{BeforeDSE}}} = \underset{n\times q}{\mathbf{G}}\ \underset{q\times p}{(\mathbf{I}\odot\mathbf{B})} + \underset{n\times K_1}{\mathbf{X}_1}\ \underset{K_1\times p}{\mathbf{B}_1} + \underset{n\times K_2}{\mathbf{X}_2}\ \underset{K_2\times p}{\mathbf{B}_2} + \underset{n\times p}{\mathbf{E}}$$

$$\underset{n\times p}{\mathbf{Y}_{\text{DSE}}} = \underset{n\times p}{\mathbf{Y}_{\text{BeforeDSE}}}\ \underset{p\times p}{(\mathbf{I}_0\odot\mathbf{B}_0)}$$

Data simulation for trans-QTL detection (design 1)

$$\underset{n\times p}{\mathbf{Y}} = \underset{n\times p}{\mathbf{Y}_{\text{BeforeDSE}}} + \underset{n\times p}{\mathbf{Y}_{\text{DSE}}}$$

$$\underset{n\times p}{\mathbf{Y}_{\text{BeforeDSE}}} = \underset{n\times q}{\mathbf{G}}\ \underset{q\times p}{(\mathbf{I}\odot\mathbf{B})} + \underset{n\times K_1}{\mathbf{X}_1}\ \underset{K_1\times p}{\mathbf{B}_1} + \underset{n\times K_2}{\mathbf{X}_2}\ \underset{K_2\times p}{\mathbf{B}_2} + \underset{n\times p}{\mathbf{E}}$$

$$\underset{n\times p}{\mathbf{Y}_{\text{DSE}}} = \underset{n\times p}{\mathbf{Y}_{\text{BeforeDSE}}}\ \underset{p\times p}{(\mathbf{I}_0\odot\mathbf{B}_0)}$$

Before Downstream effect (*cis*-QTL):

G: each element is drawn independently from Bin(2, 0.4).

I: each element is drawn independently from Ber(0.01).

B: each element is drawn independently from N(0, 4).

$\mathbf{X}_1$, $\mathbf{X}_2$: each element is drawn independently from N(0, 0.6).

$\mathbf{B}_1$, $\mathbf{B}_2$: $(\mathbf{B}_1)_{kj}$, $(\mathbf{B}_2)_{kj}$ is drawn from $N(0, \sigma_k^2)$, with the covariate-specific effect size variance $\sigma_k^2 \sim 0.8 * \left(\Gamma(2.5, 0.6)\right)^2$.

E: $\mathbf{E}_{ij}$ is drawn from $N(0, 1/\tau_j)$, with the gene-specific noise precision $\tau_j \sim \Gamma(3, 1)$.

Downstream effect (trans-QTL):

$\mathbf{I}_0$: $p\times p$ binary matrix with 3 rows containing exactly 30 ones (excluding diagonal entries).

$\mathbf{B}_0$: each element is drawn independently from N(8, 0.8) for "strong downstream effects".

$$\underset{100\times 100}{\mathbf{I}_0} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \ddots & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & \ddots & \cdots & 1 & 0 & 1 \\ \vdots & \cdots & \cdots & \mathbf{0} & \cdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \ddots & 0 & 0 \\ 1 & 1 & 0 & \cdots & 0 & \ddots & 1 \\ 1 & 0 & 0 & \cdots & 1 & 1 & 0 \end{bmatrix}$$

only 3 rows: sum = 30

Data Simulation for cis-QTL detection (design 2)

$$\underset{n\times p}{\mathbf{Y}} = \underset{n\times q\times p}{\mathbf{G}_{\text{array}}} \otimes \underset{q\times p}{(\mathbf{I}} \odot \underset{q\times p}{\mathbf{B})} + \underset{n\times K_1}{\mathbf{X}_1}\ \underset{K_1\times p}{\mathbf{B}_1} + \underset{n\times K_2}{\mathbf{X}_2}\ \underset{K_2\times p}{\mathbf{B}_2} + \underset{n\times p}{\mathbf{E}}$$

Genotype component:

$\mathbf{G}_{\text{array}}$: $\mathbf{G}_{\text{array}}[, , j]$ is an $n\times q$ real genotype matrix for the q local common SNPs for gene j.

$\mathbf{I}$: each column $\mathbf{I}_j$ is drawn independently from Mult(q; 1/q,…,1/q), or each element of $\mathbf{I}_j$ is drawn independently from Ber(1/q).

$\mathbf{B}$: each element is drawn independently from N(0, 1).

Covariate component:

$\mathbf{X}_1$, $\mathbf{X}_2$: each element is drawn independently from N(0, 1).

$\mathbf{B}_1$, $\mathbf{B}_2$: each element is drawn independently from N(0, 1) and scaled according to the specified PVE.

Noise component:

$\mathbf{E}$: each element is drawn independently from N(0, 1) and scaled according to the specified PVE.

Evaluation metrics

01

Runtime of the hidden variable inference step

We compare the runtime of 15 methods, including ideal, unadjusted methods and 13 variants of PCA, SVA, PEER, and HCP based on two simulation studies..

02

The area under the PR curve (AUPRC) of the QTL result

We calculate the area under the precision-recall curve (AUPRC).
We also compare the powers of the different methods in Simulation Design1.

03

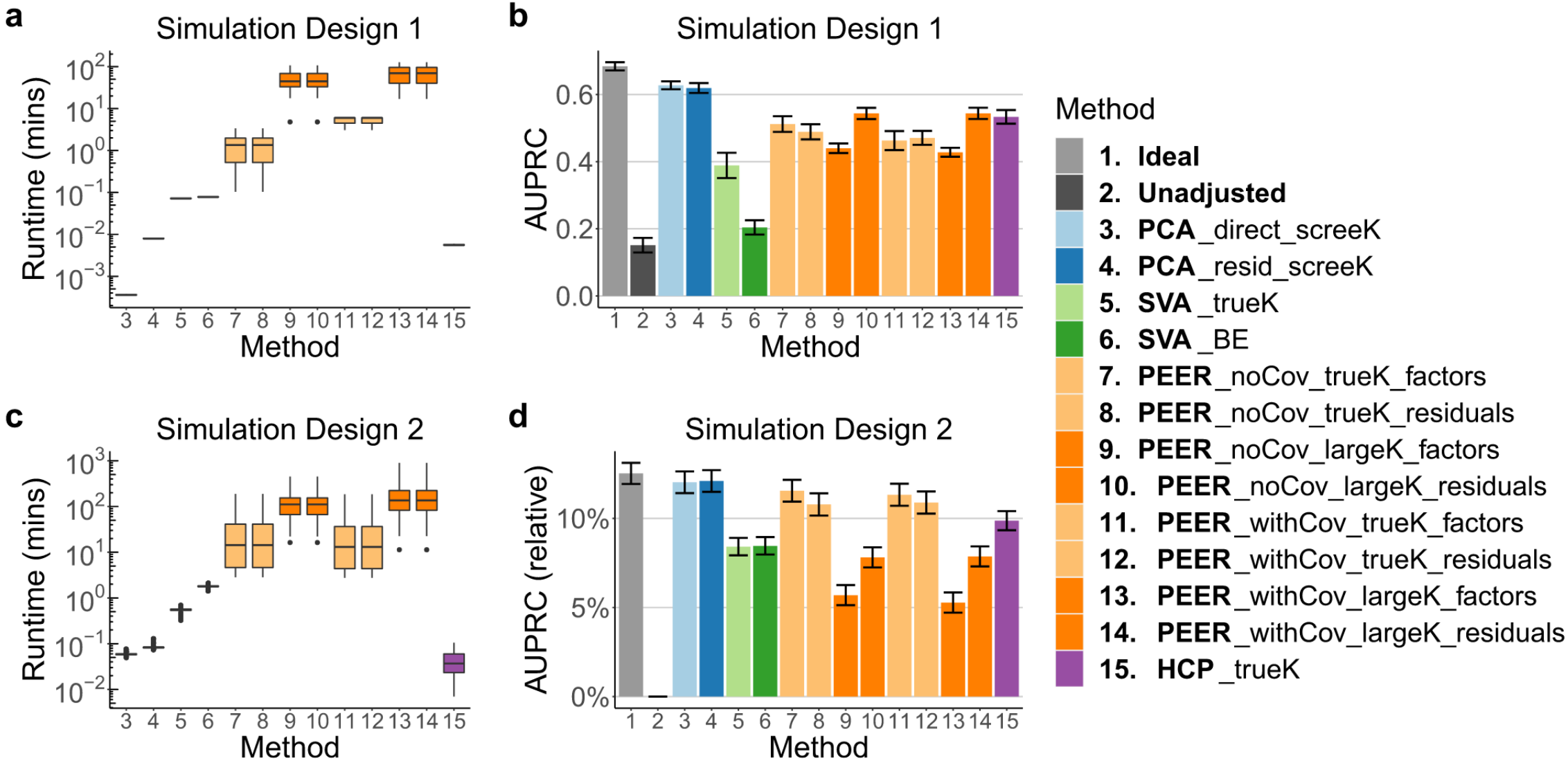
Adjusted R^2 measures

- adjusted R^2 score
(regressing each true hidden covariate against the inferred covariates)
- reverse adjusted R^2 score
(regressing each inferred covariate against the true hidden covariates)
- concordance score
(the average of the adjusted R^2 score and the reverse adjusted R^2 score)

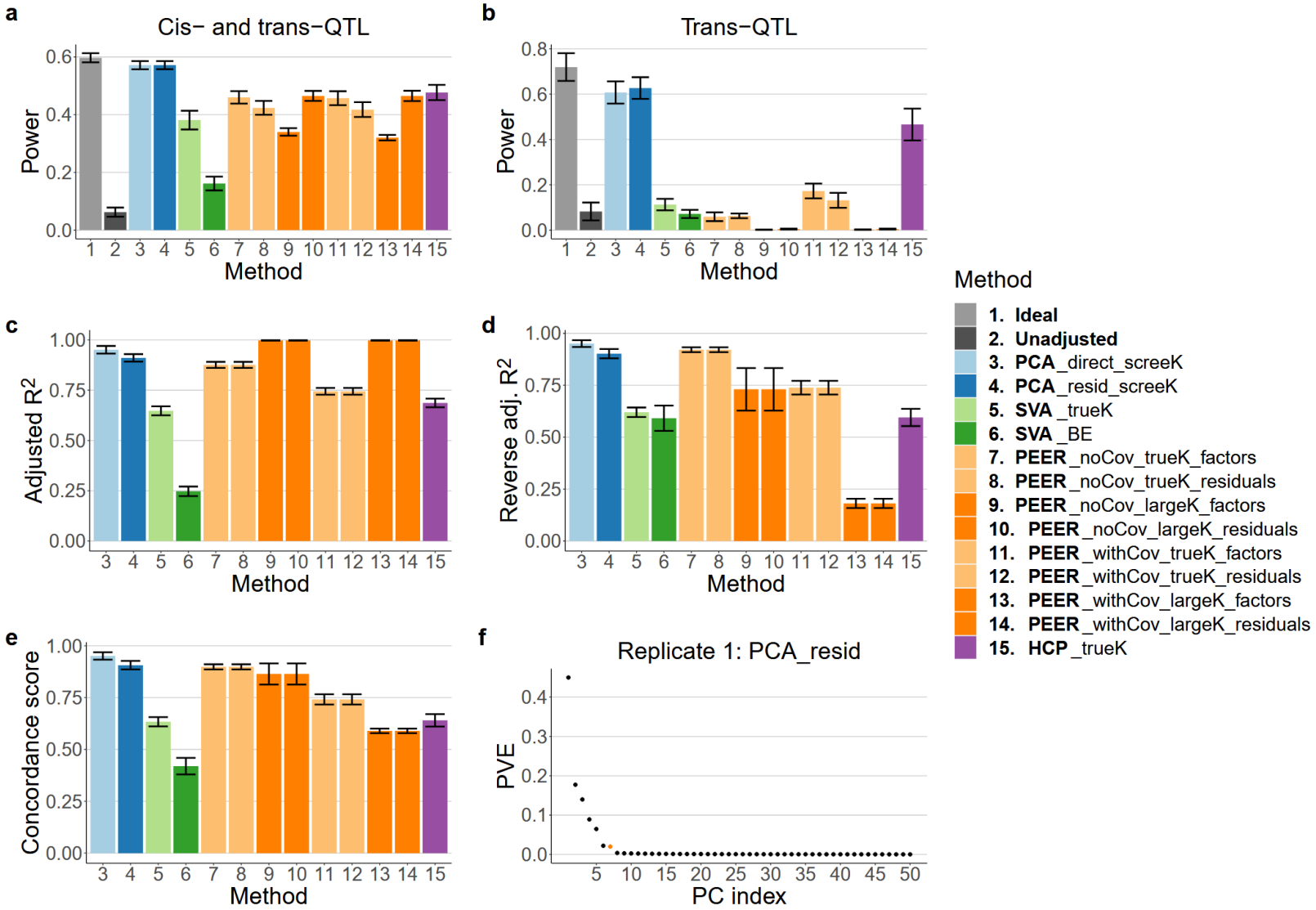
Summary of the 15 methods compared simulation studies

	Methods	Response and Covariates	Abbreviations (if selected)
1	Ideal	$\mathbf{Y}, \mathbf{X}_1 + \mathbf{X}_2$	Ideal
2	Unadjusted	$\mathbf{Y}, \mathbf{X}_1$	Unadjusted
3	**PCA**_direct_screeK	$\mathbf{Y}, \mathbf{X}_1$ (filtered) + top PCs	PCA
4	**PCA**_resid_screeK	$\mathbf{Y}, \mathbf{X}_1$+ top PCs	
5	**SVA**_trueK	$\mathbf{Y}, \mathbf{X}_1$+ SVs	
6	**SVA**_BE	$\mathbf{Y}, \mathbf{X}_1$+ SVs	SVA
7	**PEER**_noCov_trueK_factors	$\mathbf{Y}, \mathbf{X}_1$ (filtered) + PEER factors	
8	**PEER**_noCov_trueK_residuals	$\mathbf{Y}_{\text{resid}}$, NULL	
9	**PEER**_noCov_largeK_factors	$\mathbf{Y}, \mathbf{X}_1$ (filtered) + PEER factors	
10	**PEER**_noCov_largeK_residuals	$\mathbf{Y}_{\text{resid}}$, NULL	
11	**PEER**_withCov_trueK_factors	$\mathbf{Y}, \mathbf{X}_1$+ PEER factors	PEER, true K, factors
12	**PEER**_withCov_trueK_residuals	$\mathbf{Y}_{\text{resid}}$, NULL	
13	**PEER**_withCov_largeK_factors	$\mathbf{Y}, \mathbf{X}_1$+ PEER factors	
14	**PEER**_withCov_largeK_residuals	$\mathbf{Y}_{\text{resid}}$, NULL	PEER, large K, residuals
15	**HCP**_trueK	$\mathbf{Y}, \mathbf{X}_1$+ HCPs	HCP

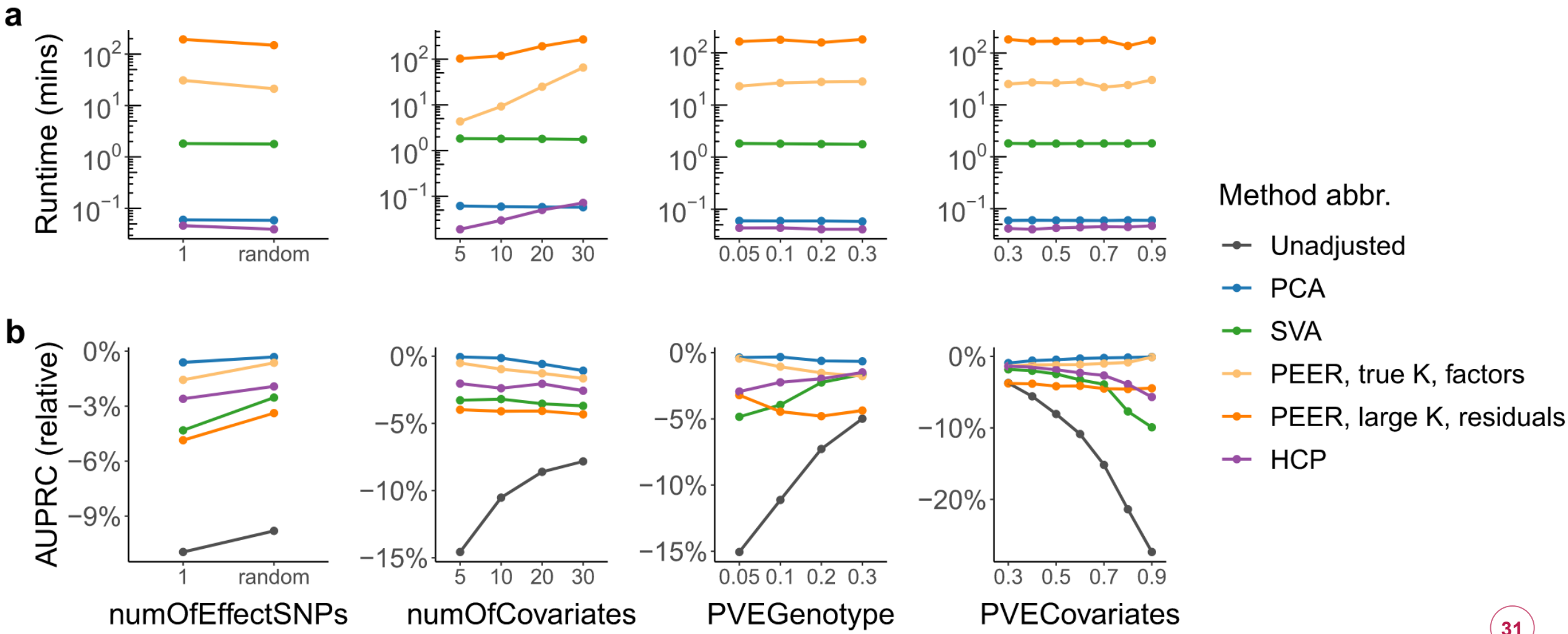
Runtime and AUPRC comparison of all 15 methods



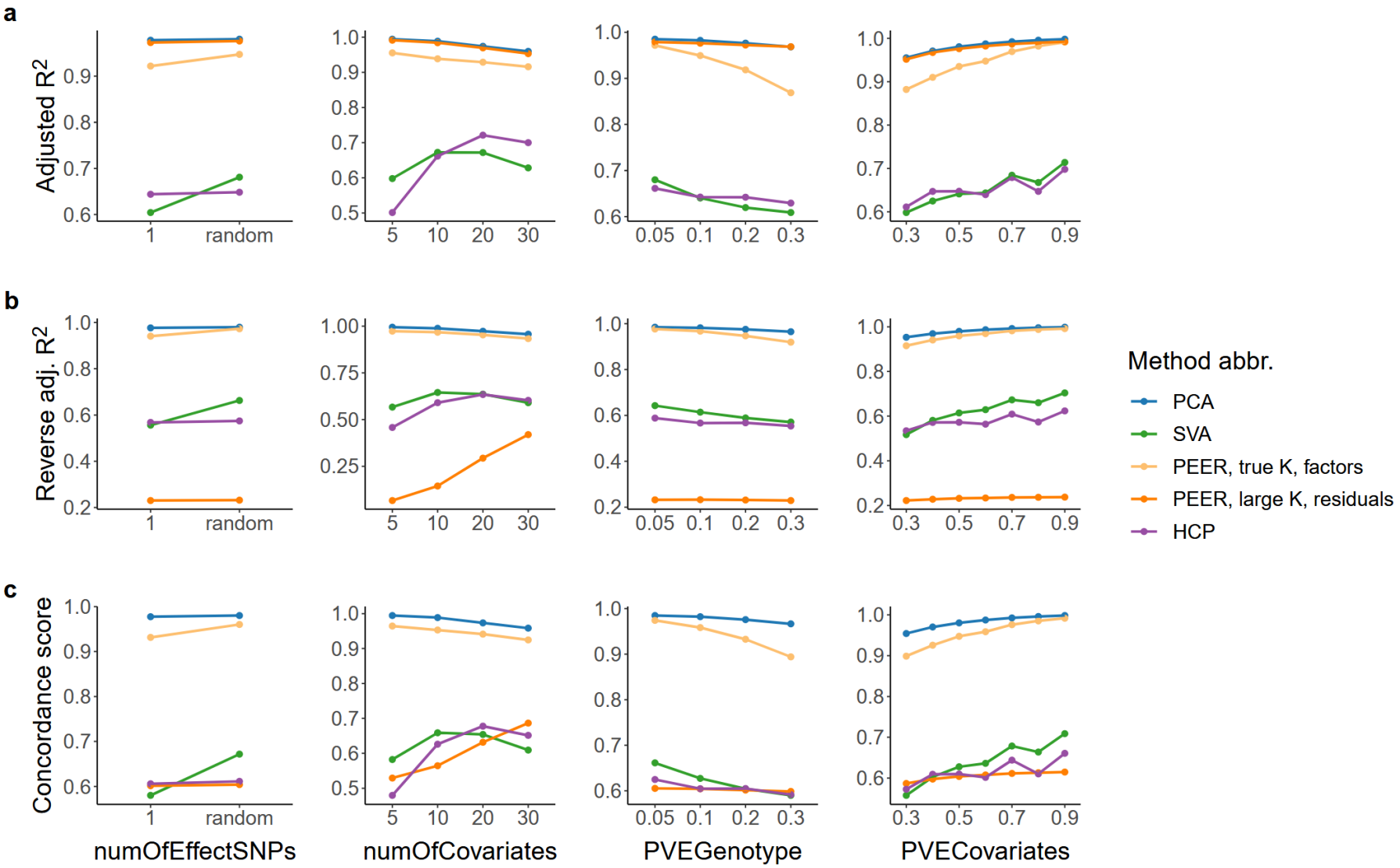
Power and adjusted R^2 measures comparison of all 15 methods in simulation design 1



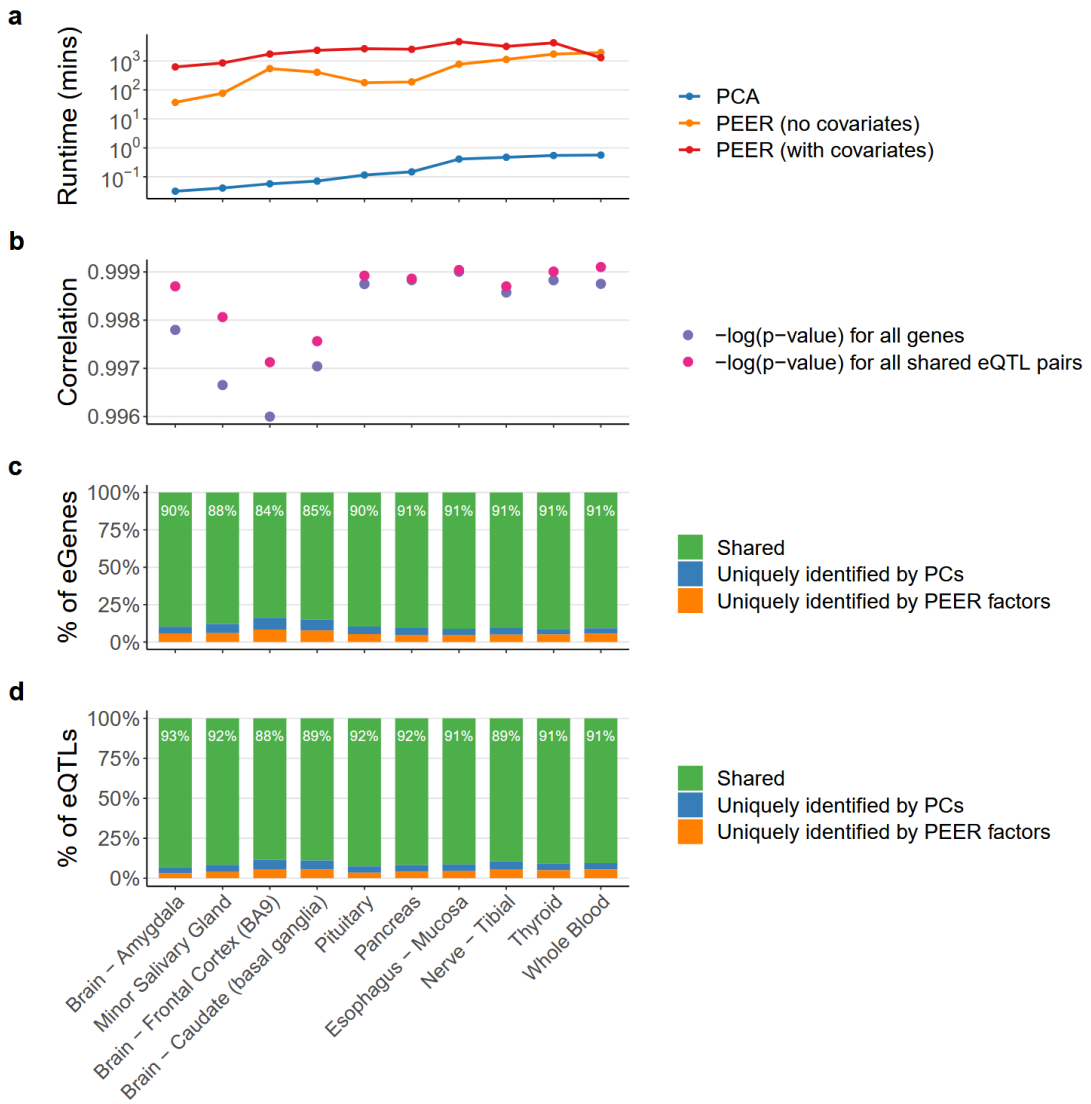
Detailed runtime and AUPRC comparison of the selected methods in simulation design 2

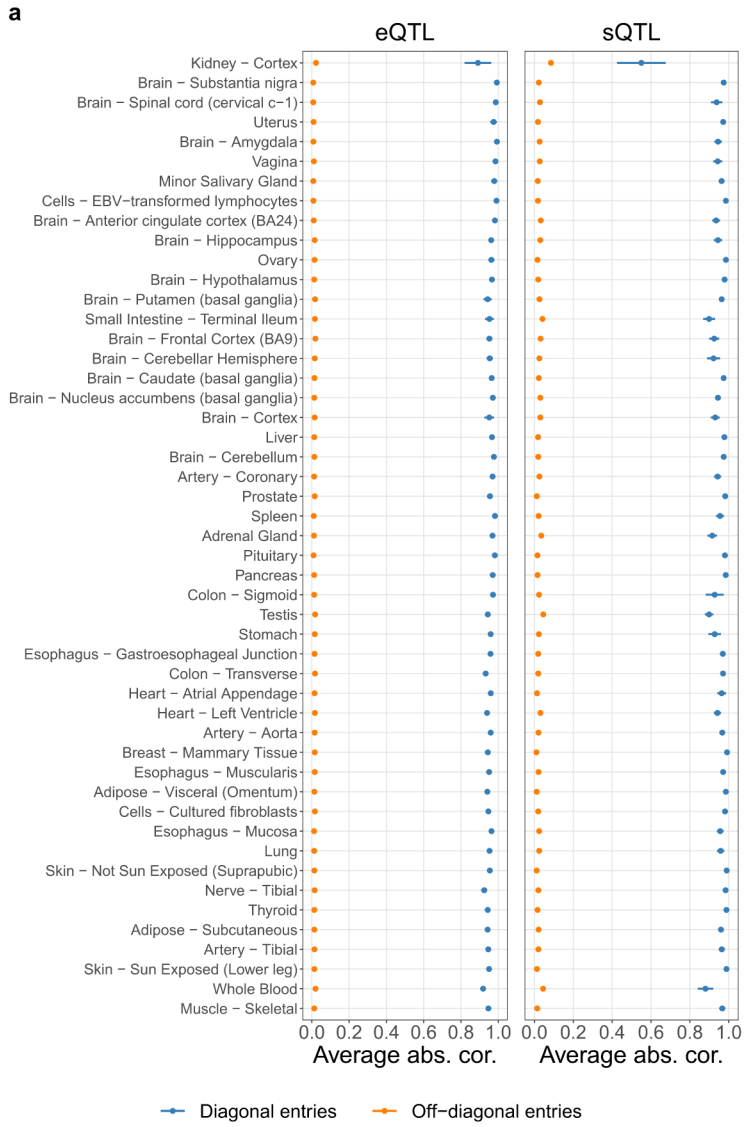


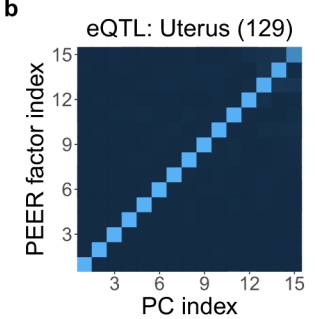
Detailed adjusted R^2 measures comparison of the selected methods in simulation design 2

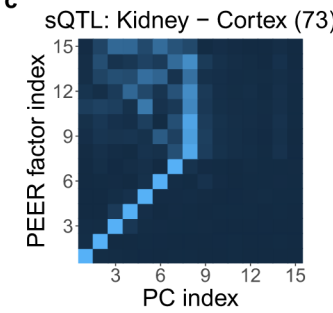


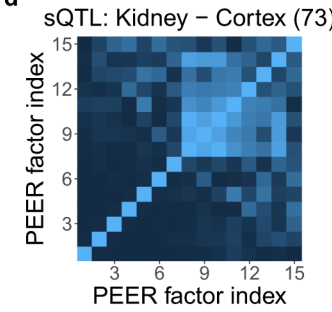
Comparison of PEER factors and PCs in GTEx eQTL and sQTL analysis











Comparison of PEER factors and PCs in GTEx eQTL and sQTL analysis

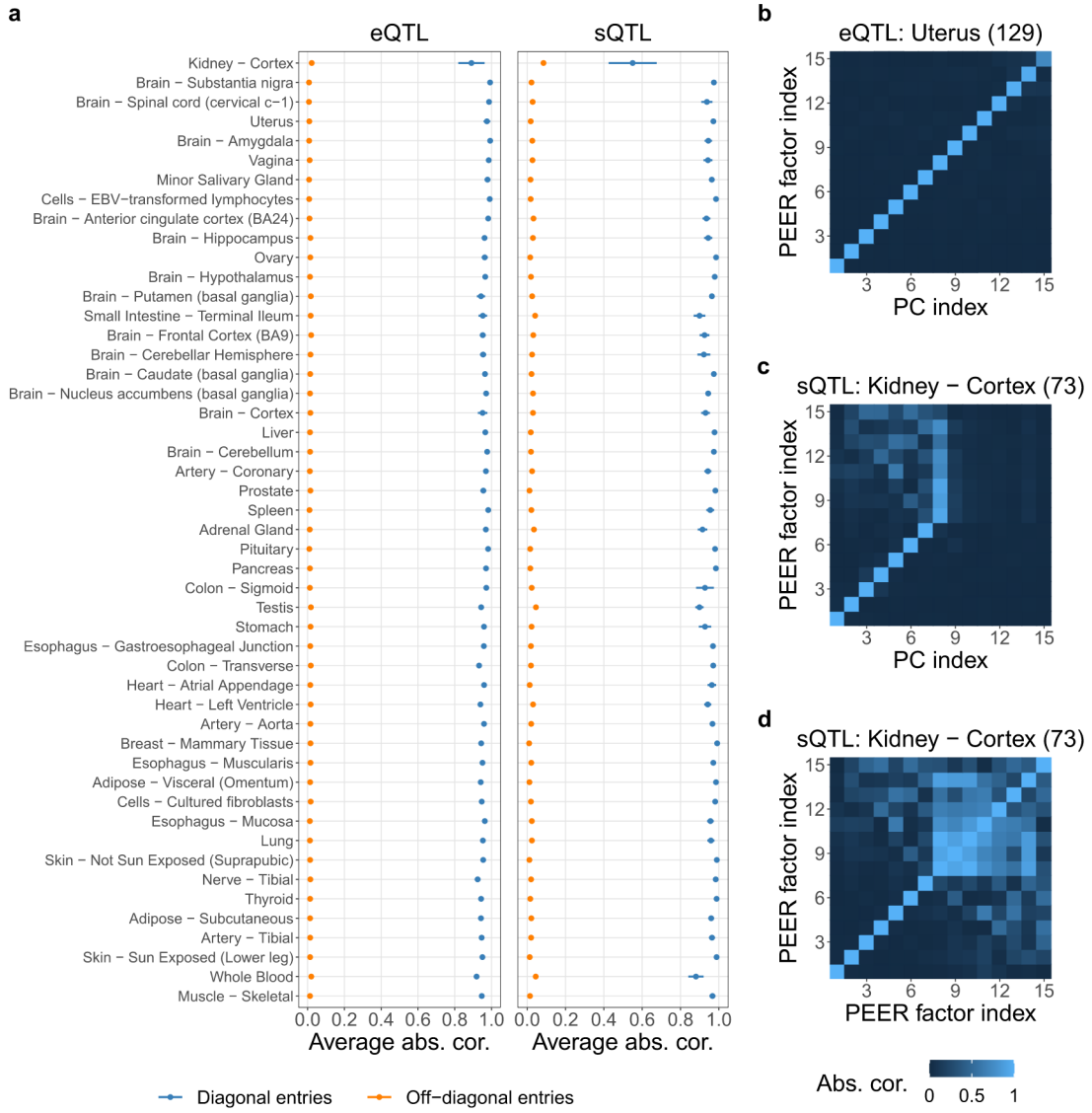
Algorithm S1: Reordering of PEER factors based on PCs (Figure 5).

Inputs:

- K PEER factors.
- K PCs.

Output: K PEER factors (reordered).

1 **for** $k \leftarrow 1$ **to** K **do**
2 | Select the PEER factor that is the most highly correlated with the kth PC from the PEER factors that have not been selected yet.
3 **end**
4 **return** the PEER factors in the order that they were selected in.

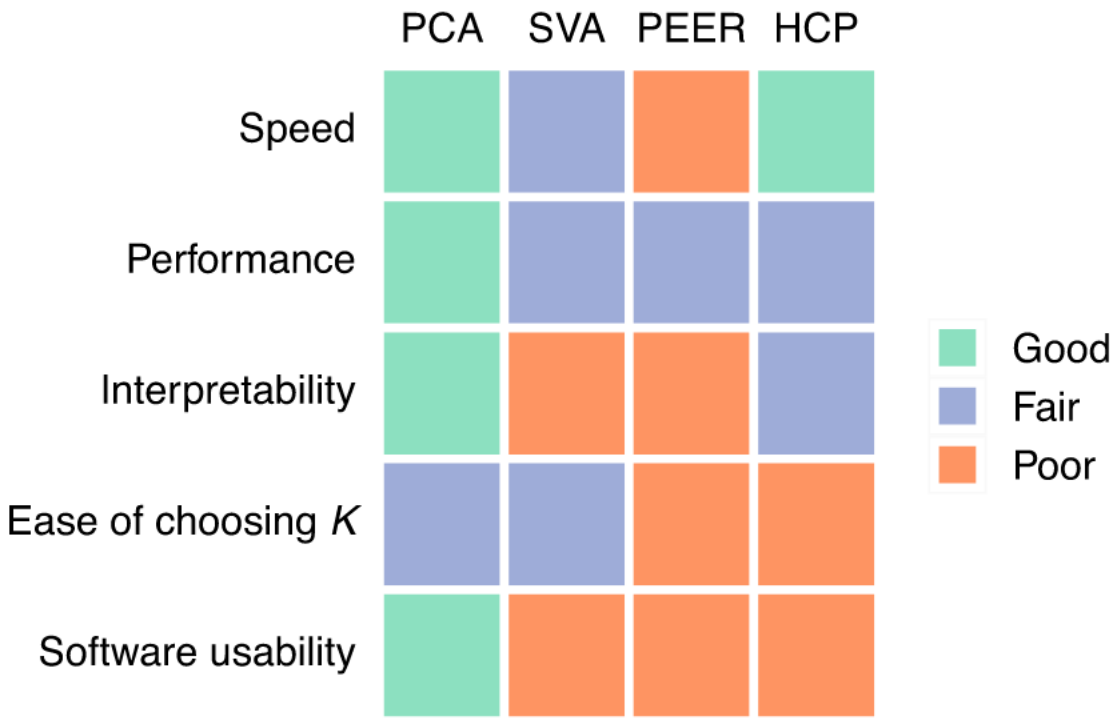


香港城市大學
City University of Hong Kong

04

Discussion

PCA outperforms other methods for the inference of the hidden variables



- PCA is faster.
- PCA is better-performing.
- PCA can be interpreted and used as both a dimension reduction and a factor discovery method.
- PCA offers convenient ways of choosing K.
- There is no consensus on how PEER, HCP and SVA should be used.

香港城市大學
City University of Hong Kong

THANK YOU

Speaker: LI Yuekai

Date: 07/11/2025

香港城市大學
City University of Hong Kong

THANK YOU

Speaker: LI Yuekai

Date: 07/11/2025