Variable selection in transcriptomics data using knockoffs in a classification framework

Julie Cartier
Supervisors : **Florian Massip** (CBIO)
Chloé-Agathe Azencott (CBIO)
Adeline Fermanian (Califrais)

Workshop on statistical inference for complex data: 12/09/2024

PhD: 14/11/2022 -

MINES PARIS | PSL

institut Curie

INSTITUT CARNOT M.I.N.E.S

Inserm

Large-scale biological data analysis comes with challenges

genes

Samples

	ENSG00000131697	ENSG00000116661	ENSG00000157330
S1_84_NS_2			
S1_85_NS_2		✖	
S1_86_NS_2			

gene expression level

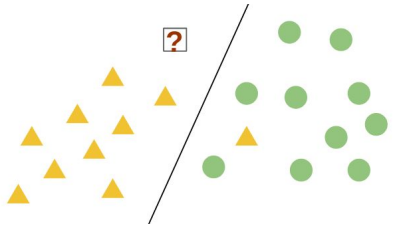
Large-scale biological data analysis comes with challenges

genes

Samples

	ENSG00000131697	ENSG00000116661	ENSG00000157330
S1_84_NS_2			
S1_85_NS_2		✖	
S1_86_NS_2			

gene expression level



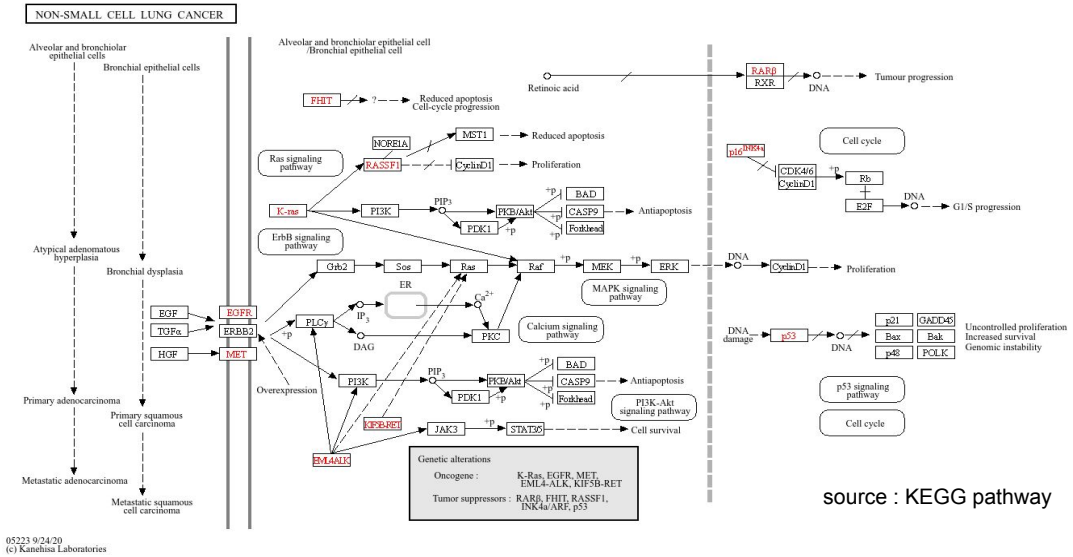
Prediction tasks

Biomarker identification

Large-scale biological data analysis comes with challenges

Biomarker identification *(Variable selection)*

- ➔ **High-dimension**: Vast amount of variables (20 000 genes), small cohorts (hundreds of patients)
- ➔ **Correlations**: Interactions between different functional units of the biological system



source : KEGG pathway

Perform variable selection with **the knockoff procedure**[1]

The Knockoff procedure[1]:

✅ Designed for correlated settings

✅ Handle high dimension (Model-X knockoff[2])

✅ Can be used with no prior knowledge about the structure of the data

- Control the expected number of false discoveries
- Already used with biological data (GWAS)[3]

Study the applicability of the knockoff procedure to **transcriptomic** data in **classification**

(1): Barber and Candes : Controlling the false discovery rate via knockoffs (2015)
(2) : Candes et al - *Panning for Gold: Model-X Knockoffs for High-dimensional Controlled Variable Selection (2017)*
(3) : Sesia et al - *False discovery rate control in genome wide association studies with population structure*

The knockoff variable selection procedure

Knockoff = False copy

X $\quad$ $\widetilde{X}$

$$X = \begin{pmatrix} X_{1,1} & \cdots & X_{1,j} & \cdots & X_{1,p} \\ X_{2,1} & \cdots & X_{2,j} & \cdots & X_{2,p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{n,1} & \cdots & X_{n,j} & \cdots & X_{n,p} \end{pmatrix} \quad \begin{matrix} \\ \\ \\ \\ X_j \end{matrix}$$

$$\widetilde{X} = \begin{pmatrix} \widetilde{X}_{1,1} & \cdots & \widetilde{X}_{1,j} & \cdots & \widetilde{X}_{1,p} \\ \widetilde{X}_{2,1} & \cdots & \widetilde{X}_{2,j} & \cdots & \widetilde{X}_{2,p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \widetilde{X}_{n,1} & \cdots & \widetilde{X}_{n,j} & \cdots & \widetilde{X}_{n,p} \end{pmatrix} \quad \begin{matrix} \\ \\ \\ \\ \widetilde{X}_j \end{matrix}$$

$$\widetilde{X} \perp\!\!\!\perp Y | X$$

Y

Outcomes

The knockoff variable selection procedure

Knockoff = False copy

X $\quad$ $\tilde{X}$

$$X = \begin{pmatrix} X_{1,1} & \cdots & X_{1,j} & \cdots & X_{1,p} \\ X_{2,1} & \cdots & X_{2,j} & \cdots & X_{2,p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{n,1} & \cdots & X_{n,j} & \cdots & X_{n,p} \end{pmatrix} \qquad \tilde{X} = \begin{pmatrix} \tilde{X}_{1,1} & \cdots & \tilde{X}_{1,j} & \cdots & \tilde{X}_{1,p} \\ \tilde{X}_{2,1} & \cdots & \tilde{X}_{2,j} & \cdots & \tilde{X}_{2,p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \tilde{X}_{n,1} & \cdots & \tilde{X}_{n,j} & \cdots & \tilde{X}_{n,p} \end{pmatrix}$$

$\tilde{X} \perp\!\!\!\perp Y | X$

$$[X, \tilde{X}] = \begin{pmatrix} X_{1,1} & \cdots & X_{1,j} & \cdots & X_{1,p} & \tilde{X}_{1,1} & \cdots & \tilde{X}_{1,j} & \cdots & \tilde{X}_{1,p} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{n,1} & \cdots & X_{n,j} & \cdots & X_{n,p} & \tilde{X}_{n,1} & \cdots & \tilde{X}_{n,j} & \cdots & \tilde{X}_{n,p} \end{pmatrix}$$

X_j $\qquad$ $\tilde{X}_j$

Effect **J** $\qquad$ Effect **J̃**

Y

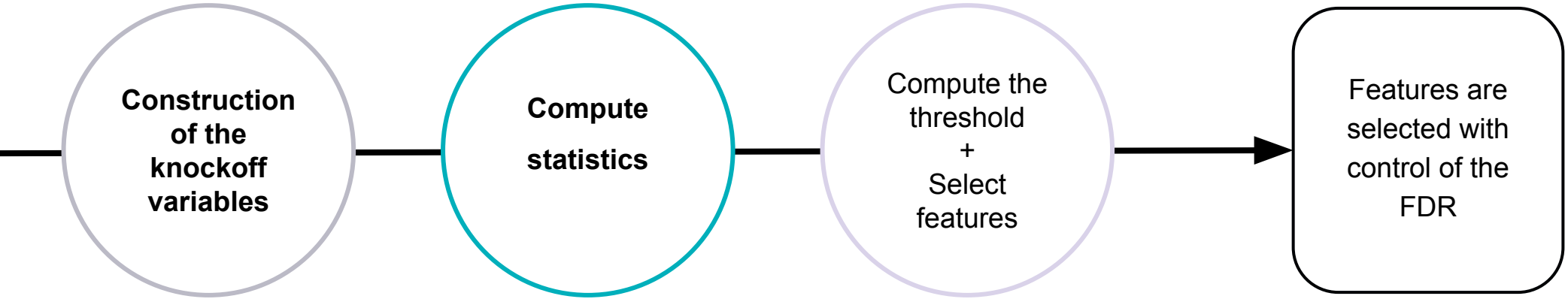
Effect **J** > Effect **J̃**:

the j^{th} variable is a relevant feature

Effect **J** ~ Effect **J̃**:

the j^{th} variable has "spurious" effect

Variable selection with Model-X knockoffs[2] procedure



➔ Adjustable to the context

➔ There are different ways to use the KO procedure

(2) : Candes et al - *Panning for Gold: Model-X Knockoffs for High-dimensional Controlled Variable Selection*

Study the applicability of the knockoff procedure to transcriptomic data in classification

→ **Performance of the method**

How to build relevant knockoff features ?

Which type of statistic performs well ?

→ **Stability**

Study the reliability of selected sets of features

→ Identified limitations of the knockoff framework

Data : Nasal transcriptomic data used to assess lung cancer risk

369 subjects (CRUKPAP[4]):

- **Nasal gene expression level : 749 risk genes**
- **4 clinical data**
  (sexe, age, smoking status, and packs years)

→

classifier:
- discriminate patients diagnosed with lung cancer from others
- Lasso penalized logistic regression

Good classification performances[4]:

- AUC-PR : **0.82**
- AUC-ROC : **0.83**

- Features are not **stably** selected

(4) : de Biase et al. - *Smoking-dependent expression alterations in the general population reveal immune impairment linked to germline variation and lung cancer risk*

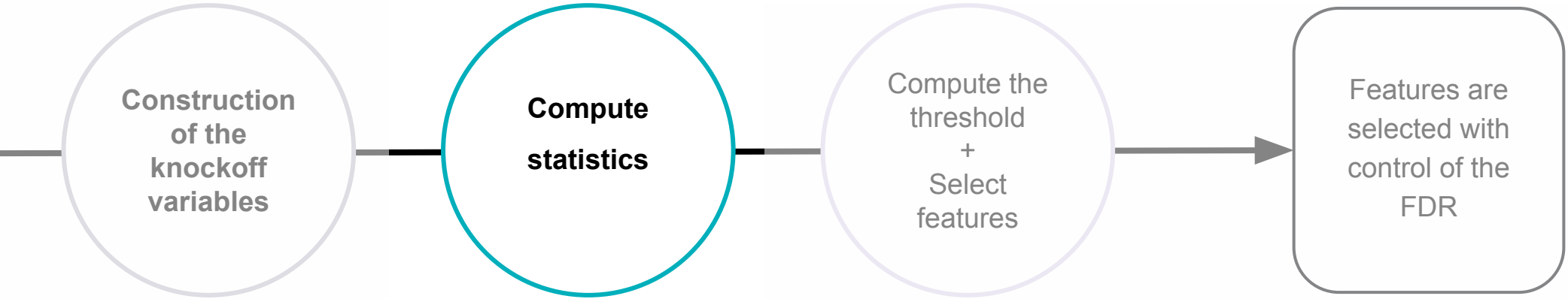
Study the applicability of the knockoff procedure to transcriptomic data in classification

→ **Performance of the method**

How to build relevant knockoff features ?

Which type of statistic performs well ?

Compute the statistics



Compute the statistics

- W_j is a function of Y and $[X, \tilde{X}]$, $\forall j \in \{1, ..., p\}$
- Large values for features that must be selected
- *Some other specific properties*

Compute statistics

$$X = \begin{pmatrix} X_{1,1} & \cdots & X_{1,j} & \cdots & X_{1,p} \\ X_{2,1} & \cdots & X_{2,j} & \cdots & X_{2,p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{n,1} & \cdots & X_{n,j} & \cdots & X_{n,p} \end{pmatrix} \quad \tilde{X} = \begin{pmatrix} \tilde{X}_{1,1} & \cdots & \tilde{X}_{1,j} & \cdots & \tilde{X}_{1,p} \\ \tilde{X}_{2,1} & \cdots & \tilde{X}_{2,j} & \cdots & \tilde{X}_{2,p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \tilde{X}_{n,1} & \cdots & \tilde{X}_{n,j} & \cdots & \tilde{X}_{n,p} \end{pmatrix}$$

X_j — Effect **J**

$\tilde{X}_j$ — Effect **J̃**

Y

Effect **J** > Effect **J̃** or Effect **J** ~ Effect **J̃** ?

Test statistics benchmark

$$[X, \tilde{X}] = \begin{pmatrix} X_{1,1} & \cdots & X_{1,j} & \cdots & X_{1,p} & \tilde{X}_{1,1} & \cdots & \tilde{X}_{1,j} & \cdots & \tilde{X}_{1,p} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{n,1} & \cdots & X_{n,j} & \cdots & X_{n,p} & \tilde{X}_{n,1} & \cdots & \tilde{X}_{n,j} & \cdots & \tilde{X}_{n,p} \end{pmatrix}$$

Variable Importance (VI) - based methods:

- Fit a ML model to $[X, \tilde{X}], Y$.
- Get $VI(j)$ and $VI(j+p), \forall j \in \{1, ..., p\}$
- $W_j = VI(j) - VI(j+p)$

Test statistics benchmark

$$[X, \tilde{X}] = \begin{pmatrix} X_{1,1} & \cdots & X_{1,j} & \cdots & X_{1,p} & \tilde{X}_{1,1} & \cdots & \tilde{X}_{1,j} & \cdots & \tilde{X}_{1,p} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{n,1} & \cdots & X_{n,j} & \cdots & X_{n,p} & \tilde{X}_{n,1} & \cdots & \tilde{X}_{n,j} & \cdots & \tilde{X}_{n,p} \end{pmatrix}$$

Variable Importance (VI) - based methods:

- Fit a ML model to $[X, \tilde{X}], Y$.
- Get $VI(j)$ and $VI(j+p), \forall j \in \{1, ..., p\}$
- $W_j = VI(j) - VI(j+p)$

- LASSO
- Elastic-net (EN)
- Random Forest (RF)
- Boosted Tree
- Deep learning

Data simulation

➤ Simulated outcomes from a given matrix $X \in \mathbb{R}^{369 \times 749}$:

Use **10** non null features

(X_{real} : the real transcriptomic matrix)

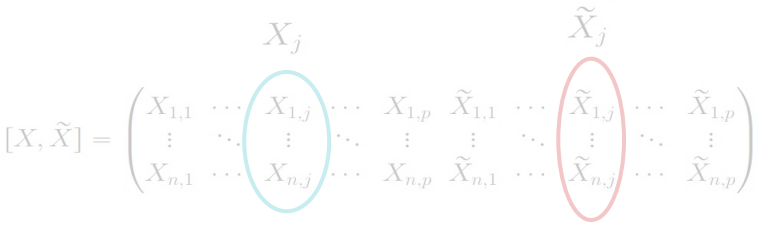
Classification:

$$Y = \mathbb{1}_{\left(\frac{1}{1+e^{-X_n\beta}} > 0.5\right)} = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}$$

Linear setting:

$$X_n\beta = X_1\beta_1 + X_2\beta_2 + \cdots + X_p\beta_p$$

Test statistics benchmark



Variable Importance (VI) - based methods:

- Fit a ML model to $[X, \widetilde{X}], Y$.
- Get $VI(j)$ and $VI(j+p), \forall j \in \{1, ..., p\}$
- $W_j = VI(j) - VI(j+p)$

- LASSO
- Elastic-net (EN)
- Random Forest (RF)
- Boosted Tree
- Deep learning

LASSO Coefficient-Difference[(2)] (LCD)

\+ Other methods

(2) : Candes et al - *Panning for Gold: Model-X Knockoffs for High-dimensional Controlled Variable Selection*

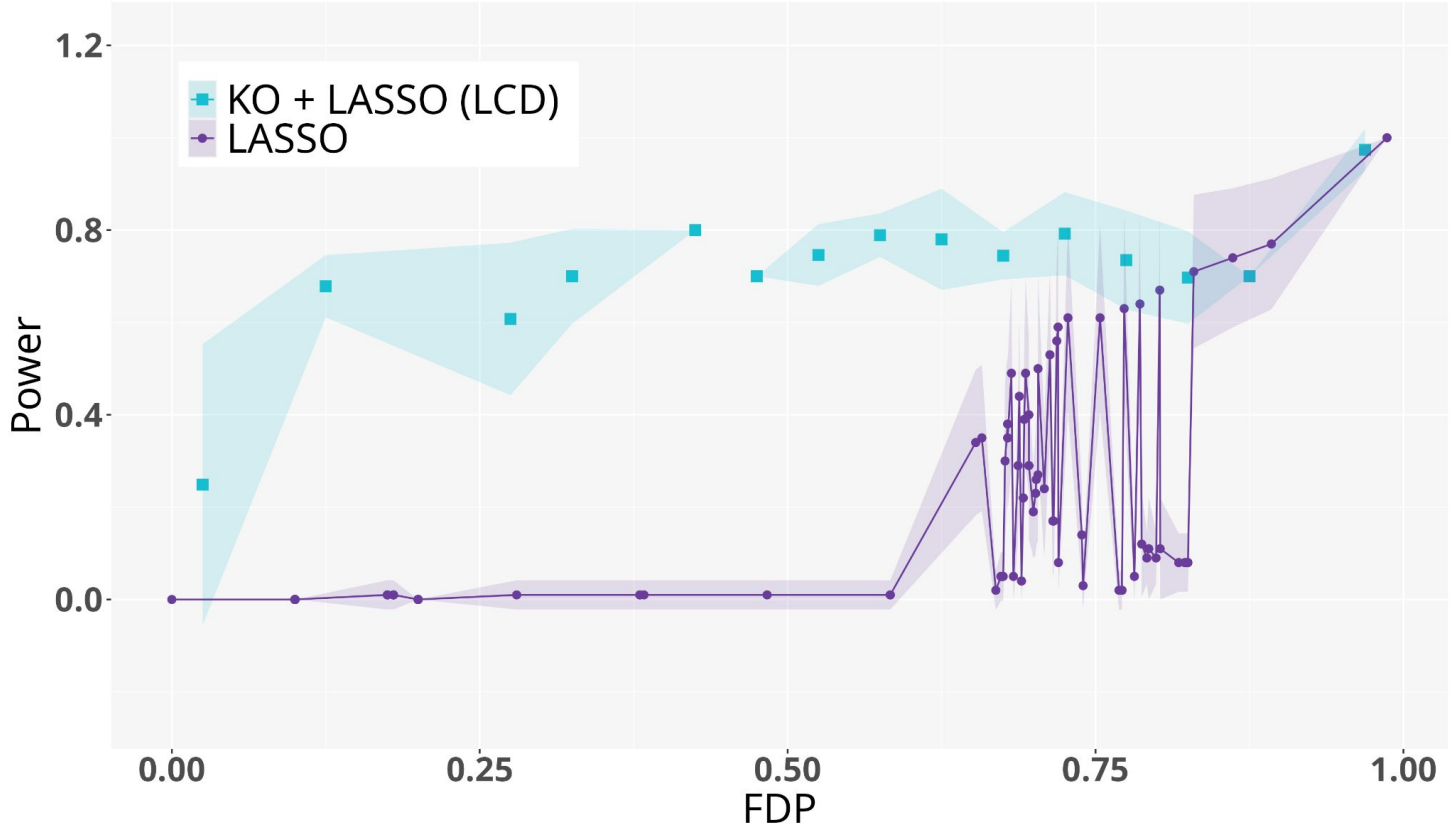
LCD based knockoff improves variable selection

Power:

$$\frac{\text{\# True discoveries}}{\text{\# True features}}$$

FDP:

$$\frac{\text{\# False discoveries}}{\text{\# Discoveries}}$$



➔ The **LCD** knockoff is more powerful than the lasso penalized logistic regression for reasonable values of FDR/FDP.

- The **LCD** based knockoff is more powerful than the lasso penalized logistic regression for small numbers of false discoveries

Study the applicability of the knockoff procedure to transcriptomic data in classification

→ **Stability**

Study the reliability of selected sets of features

(+ assess method's reliability with no ground truth)

How to get selection frequency

Number of features

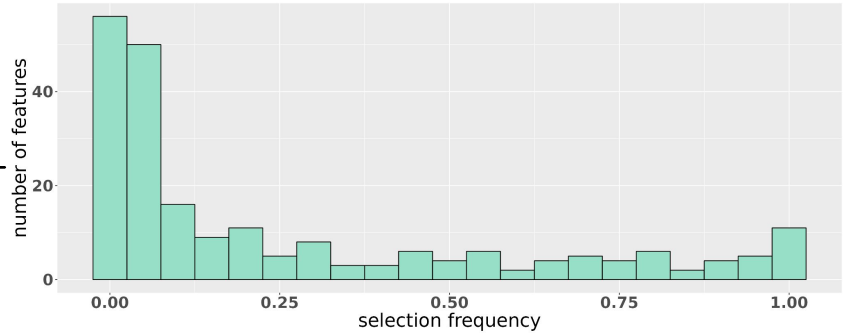
$$\begin{pmatrix} 0 & \cdots & 1 \\ 0 & & 0 \\ \vdots & \ddots & \vdots \\ 0 & & 1 \\ 1 & \cdots & 0 \end{pmatrix}$$

100 (10 x 10 fold CV with 369 patients)

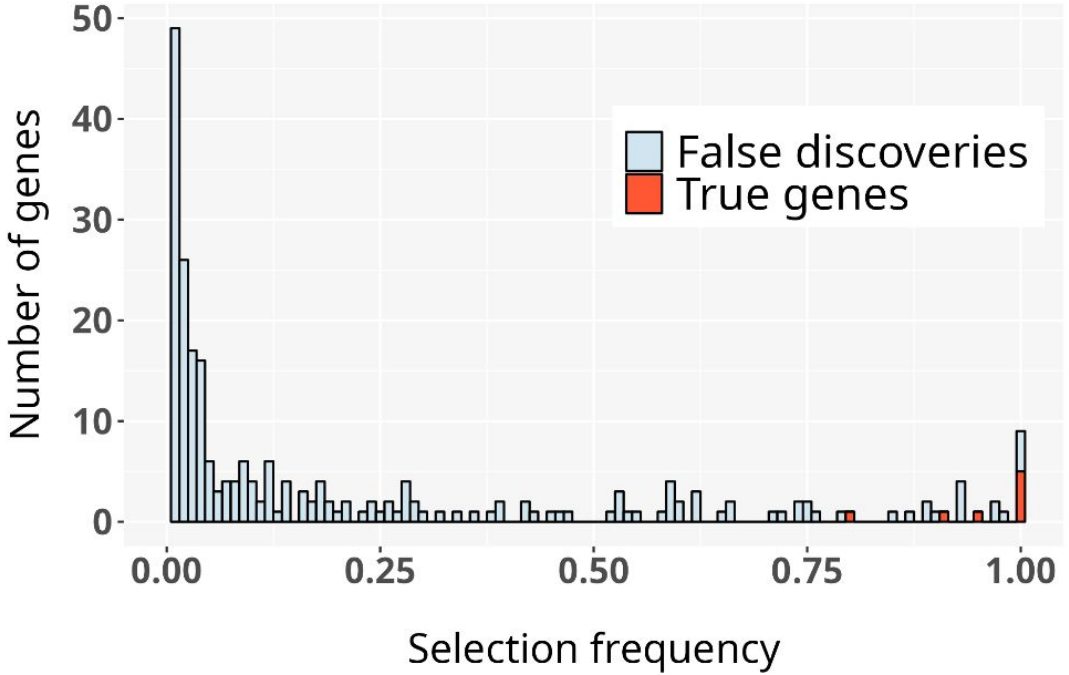
gene 1 selection frequency

.
.
.

gene p selection frequency



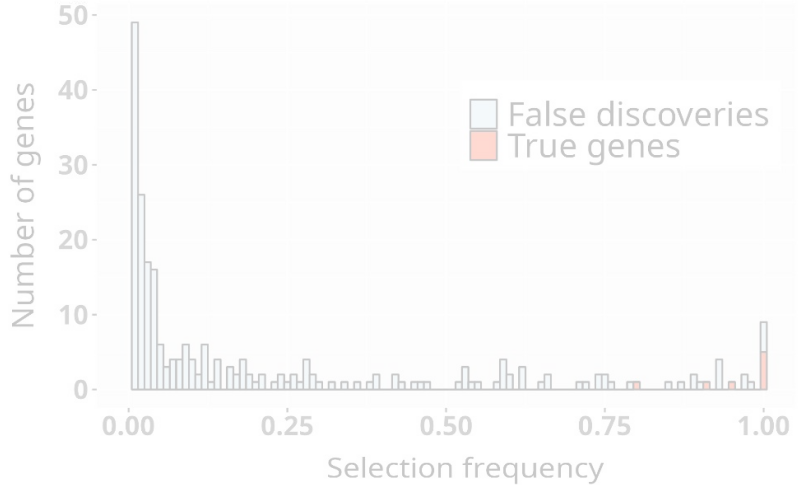
Lasso systematically selects both important and noisy features

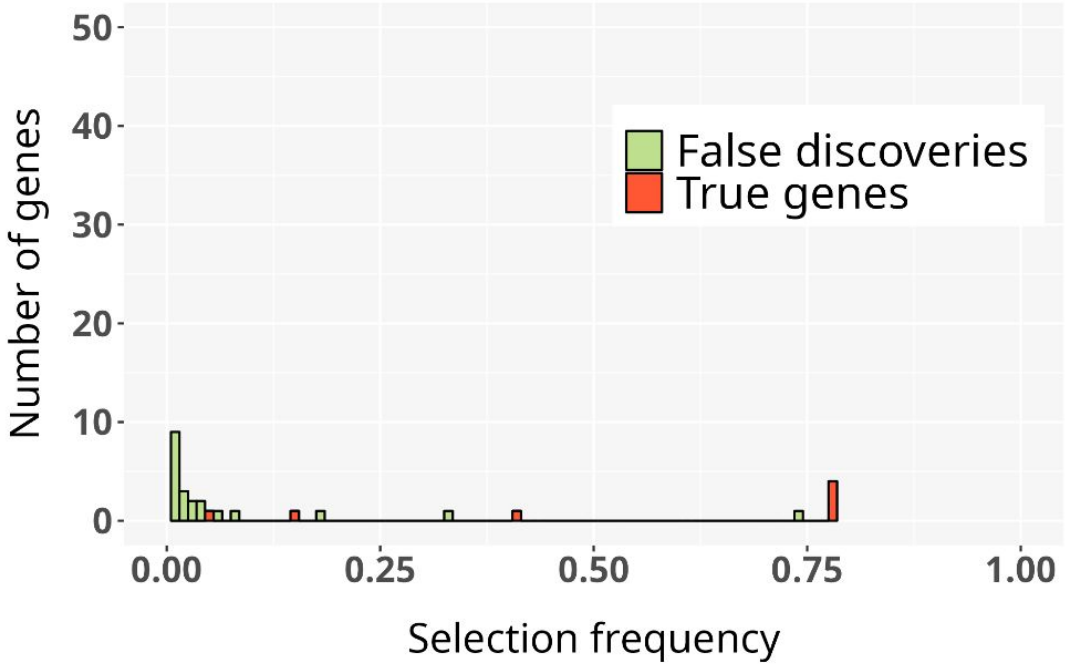


Features selected in the **lasso penalized logistic regression**

X	real matrix
Y	**Linear simulation**

The KO framework removes false positive genes from the set of most selected genes…





X	real matrix
Y	**Linear simulation**

Features selected with the **knockoff** framework.

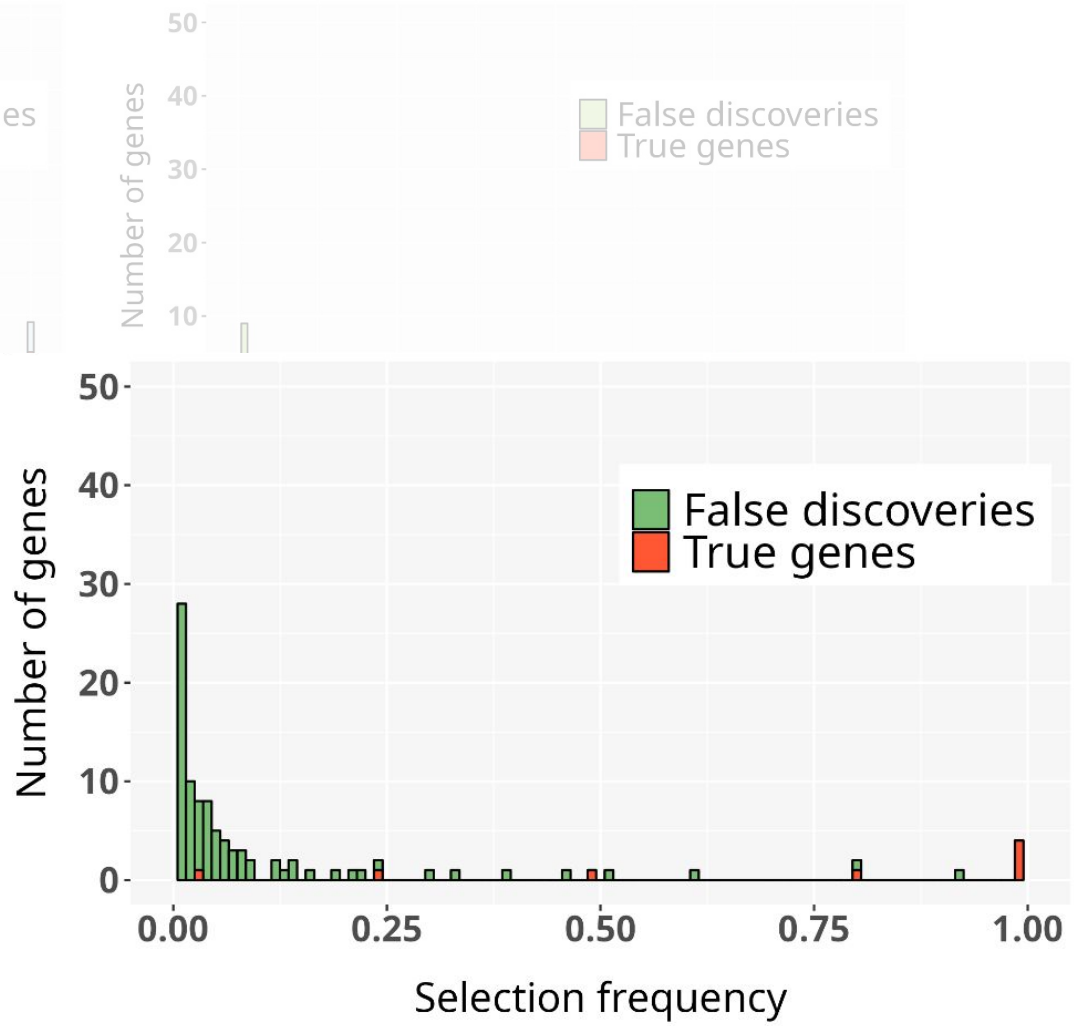
Target FDR level (q) = **0.2**

The KO framework removes false positive genes from the set of most selected genes…

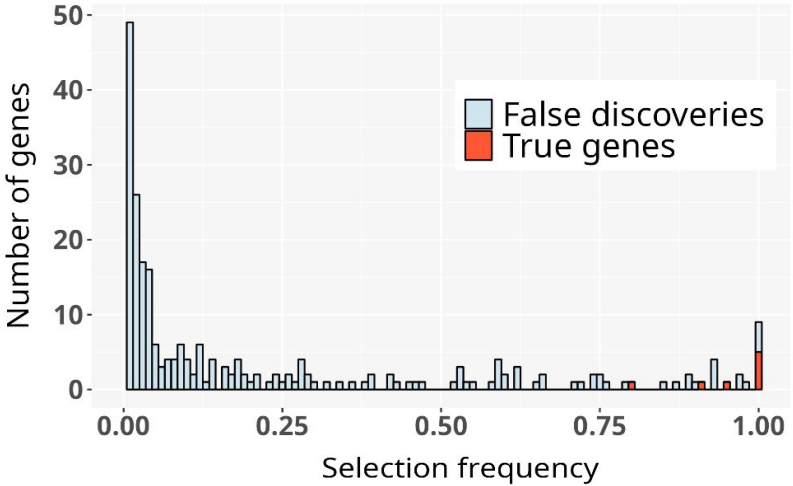
Features selected with the **knockoff** framework.

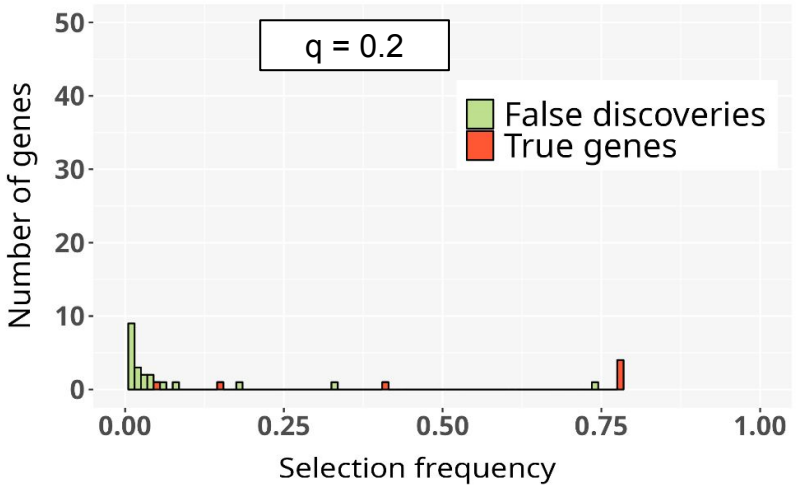
Target FDR level (q) = **0.5**

X	real matrix
Y	**Linear simulation**

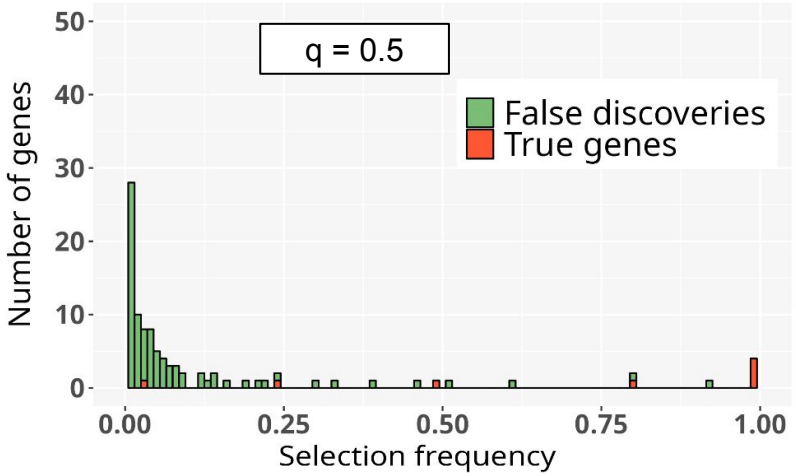


The KO framework removes false positive genes from the set of most selected genes…



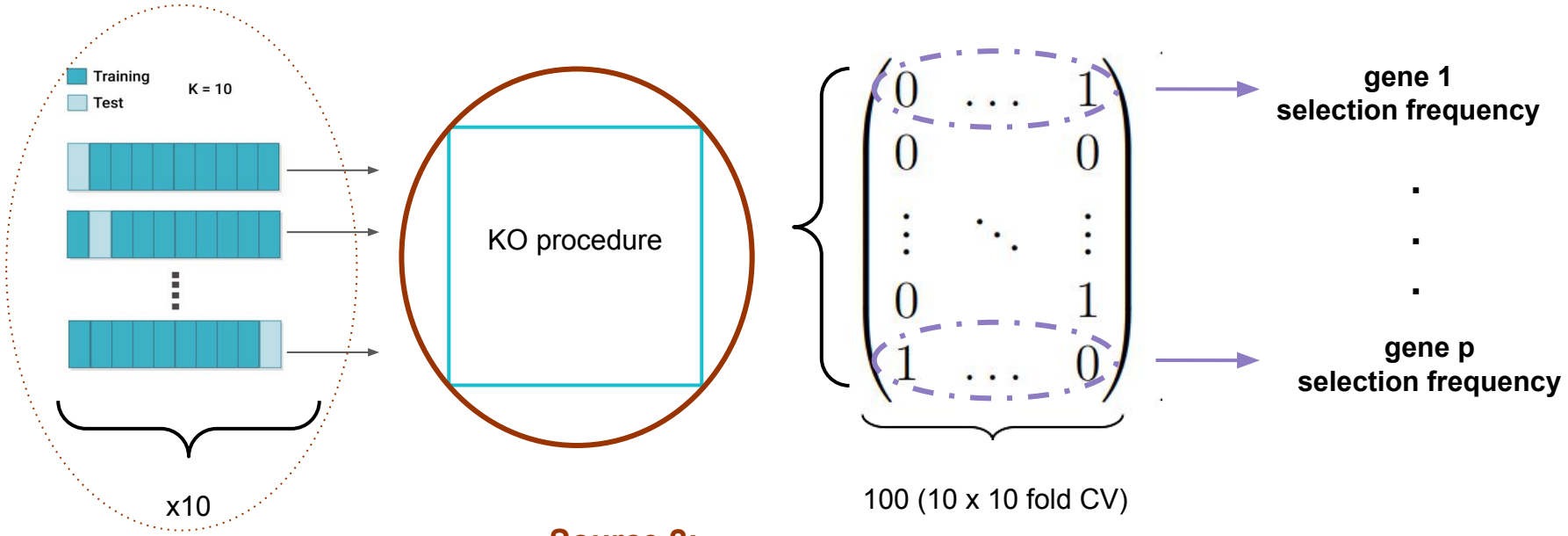


X	real matrix
Y	**Linear simulation**



… features are still not stably selected

Additional source of instability with the knockoff framework



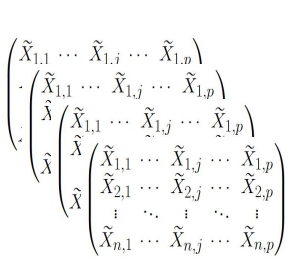
Source 1

Source 2:

- Methods to build knockoffs are **stochastic**

Knockoff aggregation

Run the knockoff procedure multiple times (in parallel):



Statistics for all features

(at each run)

Selected sets of genes

$$W_{\cdot}$$
$$W_{\cdot}$$
$$W_{\cdot}$$
$$W_j$$

$$\hat{\mathcal{S}} = \{j : W_i \geqslant \tau\}$$
$$\hat{\mathcal{S}} = \{j : W_i \geqslant \tau\}$$
$$\hat{\mathcal{S}} = \{j : W_i \geqslant \tau\}$$
$$\hat{\mathcal{S}} = \{j : W_j \geqslant \tau\}$$

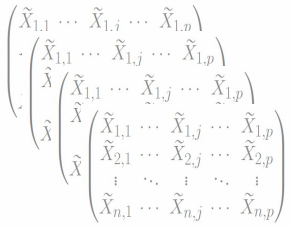
➔ Combine the results *(e.g. frequency, harmonic mean, …)* into new statistics *(e.g. p-value, π-statistics,...).*

➔ Perform inference from this new statistics

➔ More robust variable selection *(with control of the number of false discoveries)*

Knockoff aggregation

Run the knockoff procedure multiple times (in parallel):



Statistics for all features

(at each run)

Selected sets of genes

$$W_j$$

$$\hat{\mathcal{S}} = \{j : W_j \geqslant \tau\}$$

- → Combine the results *(e.g. frequency, harmonic mean, …)* into new statistics *(e.g. p-value, π-statistics,...).*
- → Perform inference from this new statistics
- → More robust variable selection *(with control of the number of false discoveries)*

KOPI

False Discovery Proportion control for aggregated Knockoffs

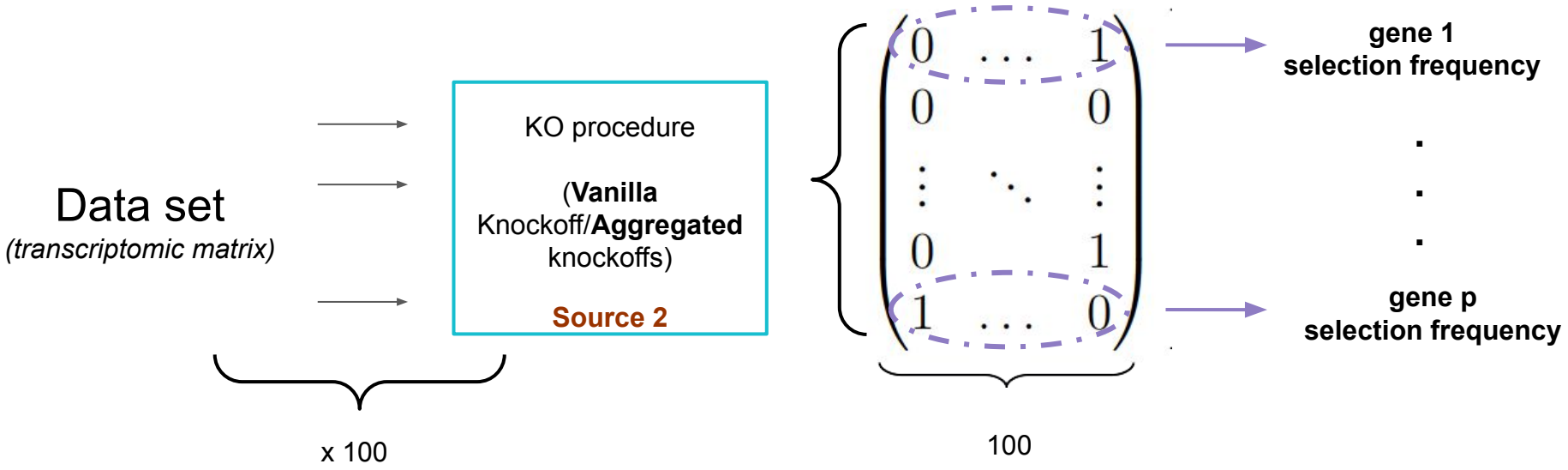
Alexandre Blain
INRIA
Université Paris-Saclay
alexandre.blain@inria.fr

Bertrand Thirion
INRIA
CEA
bertrand.thirion@inria.fr

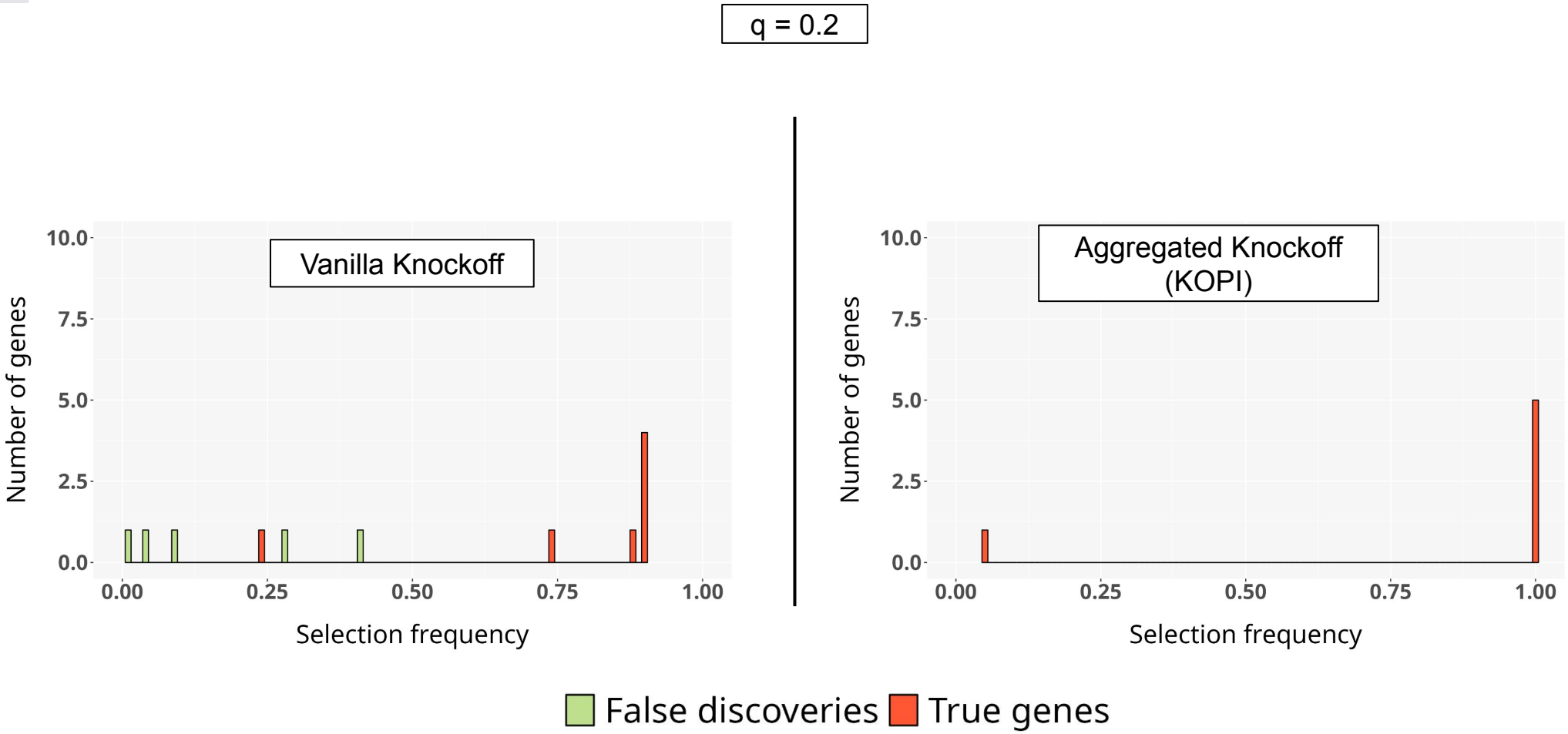
Olivier Grisel
INRIA
olivier.grisel@inria.fr

Pierre Neuvial
Institut de Mathématiques de Toulouse
Université de Toulouse
pierre.neuvial@math.univ-toulouse.fr

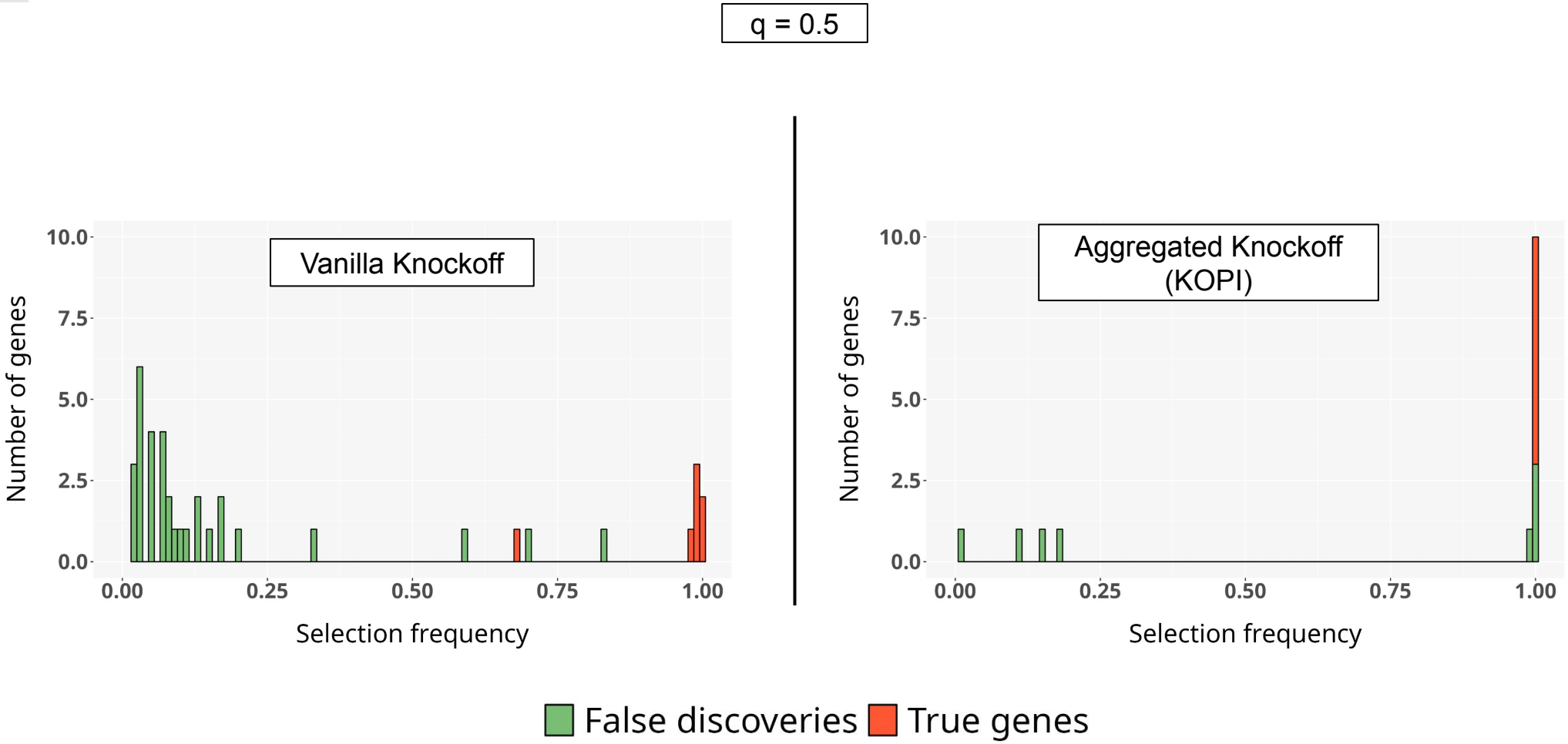
Evaluate aggregation knockoff efficiency



The aggregation of knockoffs tends to remove alternatively selected genes



The aggregation of knockoffs tends to remove alternatively selected genes



- The **LCD** based knockoff is more powerful than the lasso penalized logistic regression for small numbers of false discoveries
- The KO procedure does not improve the stability of features selection in our setting
  - Aggregating knockoffs eliminates some of the effect of knockoff stochasticity

- The **LCD** based knockoff is more powerful than the lasso penalized logistic regression for small numbers of false discoveries
- The KO procedure does not improve the stability of features selection in our setting
  - Aggregating knockoffs eliminates some of the effect of knockoff stochasticity

→ **Identified limitations of the knockoff framework**

- The knockoff procedure fails with interacting variables
- The power decreases with the number of non null features

- The **LCD** based knockoff is more powerful than the lasso penalized logistic regression for small numbers of false discoveries
- The KO procedure does not improve the stability of features selection in our setting
  - Aggregating knockoffs eliminates some of the effect of knockoff stochasticity
- Identified limitations of the knockoff procedure
  - Powerless in the interaction setting
  - Decrease in power as the number of non-null features increases
- The knockoff procedure does not select any genes with real outcomes

Group knockoffs

Why ?

- Improve the interaction setting
- Perform selection at a higher functional level (more robust)
- Integrate more features

Group knockoffs

Why ?

- Improve the interaction setting
- Perform selection at a higher functional level (more robust)
- Integrate more features

How ?

- How to make groups ?
  - e.g.: transcription factors, biological pathways, ….
- How to build knockoffs ?
  - Either use previous methods or use group's structure instead of correlations between features
- At which level does the statistic do the selection (feature or group) ?
  - group level

Acknowledgements

- **Florian MASSIP**
- **Adeline FERMANIAN**
- **Chloé-Agathe AZENCOTT**
- Johanna LAGOAS

CBIO members

Thank you for your attention !

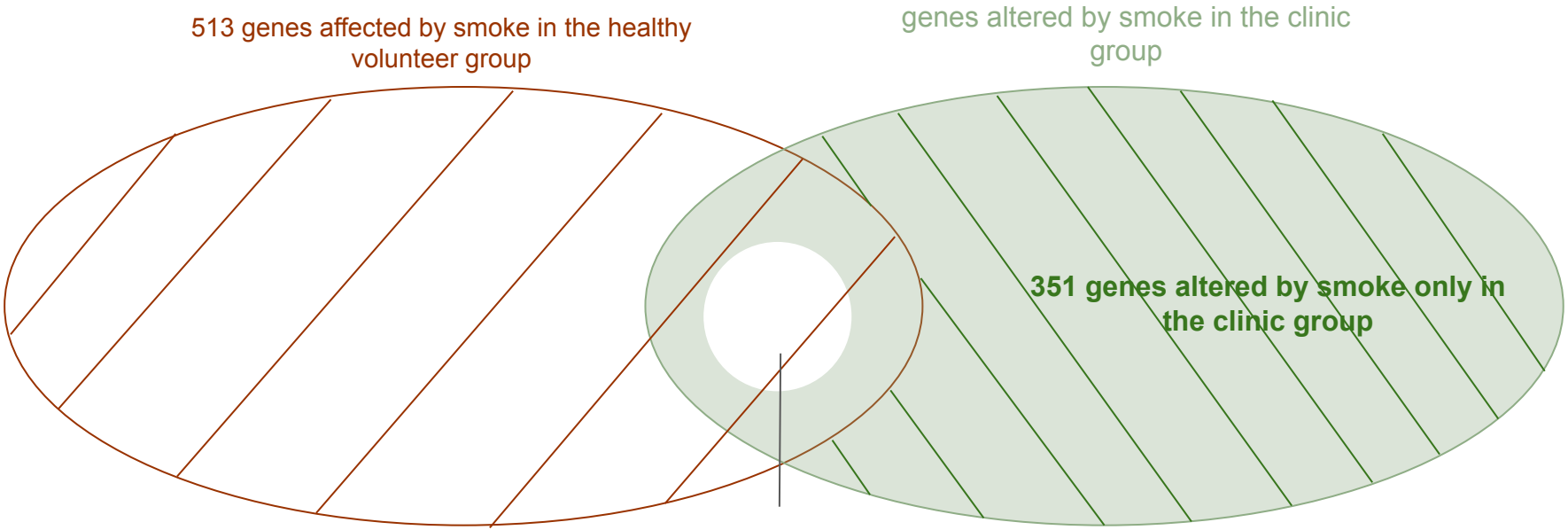
MINES PARIS | PSL

institut Curie

INSTITUT CARNOT M.I.N.E.S

Inserm

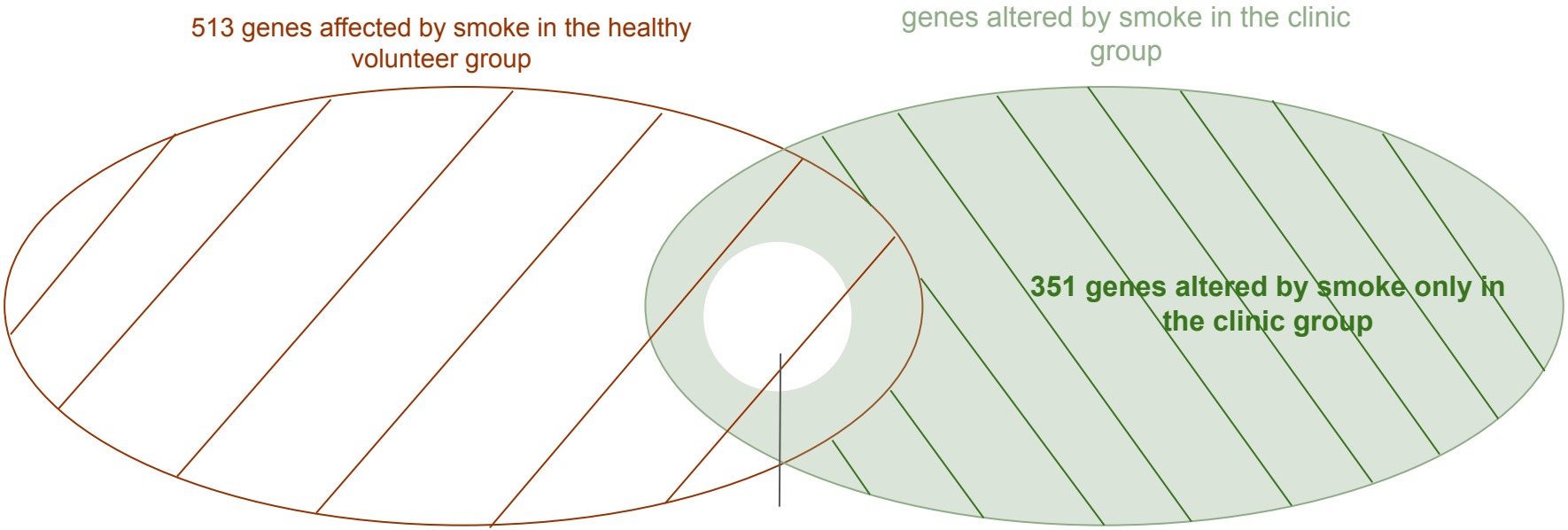
749 risk genes



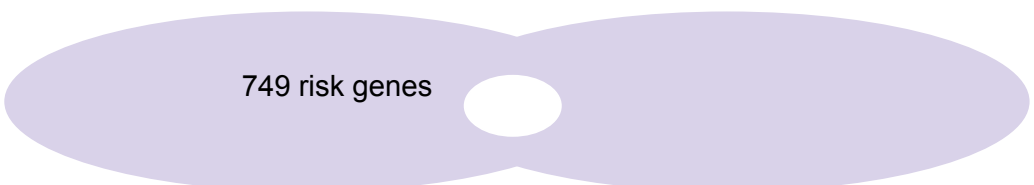
-115 shared DE genes + same gene expression dynamic/behavior

(3) : de Biase et al. - Smoking-dependent expression alterations in the general population reveal immune impairment linked to germline variation and lung cancer risk

749 risk genes



-115 shared DE genes + same gene expression dynamic/behavior



(3) : de Biase et al. - Smoking-dependent expression alterations in the general population reveal immune impairment linked to germline variation and lung cancer risk

Generation of the knockoff (KO) : LSCIP algorithm

Linear sequential conditional independent pairs[4]
(LSCIP)

- Does not use the covariance matrix
- no assumption on features distribution

- computationally more expensive

LSCIP pseudo-algorithm :

- $j = 1$, for all $j \in \{1, ..., p\}$:
  - Fit a lasso model on (X_{-j}) with X_j as outcomes.
  - Compute residuals $\epsilon = (\epsilon_1, ..., \epsilon_p)$, $\epsilon = X_j - \hat{X}_j$ where $\hat{X}_j$ is the predicted value of X_j with the regression model
  - Permute the residuals vectors randomly. Let's denote ρ_p the permutation of $\{1, ..., p\}$. compute $\tilde{X}_j = \hat{X}_j + \epsilon_{\rho_p(j)}$
- Return $\tilde{X}$

---

(4) : Blain et al - *False discovery proportion control for aggregated knockoffs*

Variable selection with Model-X knockoffs[2] procedure

Construction of the knockoff variables:

$$\widetilde{X}$$

Machine learning model: with X, $\widetilde{X}$ as features and Y as outcomes

Compute statistics:

$$W_j$$

$$\forall j \in \{1, ..., p\}$$

Select features with statistics above a given threshold:

$$\hat{\mathcal{S}} = \{j : W_j \geqslant \tau\}$$

τ is a function of:

- a target FDR level q
- W_j

$$\tau := \min\left\{t > 0 : \frac{1 + \#\{j : W_j \leqslant -t\}}{\#\{j : W_j \geqslant t\}} \leqslant q\right\}$$

(2) : Candes et al - *Panning for Gold: Model-X Knockoffs for High-dimensional Controlled Variable Selection*

Variable selection with Model-X knockoffs[2] procedure

Construction of the knockoff variables:

$$\widetilde{X}$$

Machine learning model: with X, $\widetilde{X}$ as features and Y as outcomes

Compute statistics:

$$W_j$$

$$\forall j \in \{1, ..., p\}$$

Select features with statistics above a given threshold:

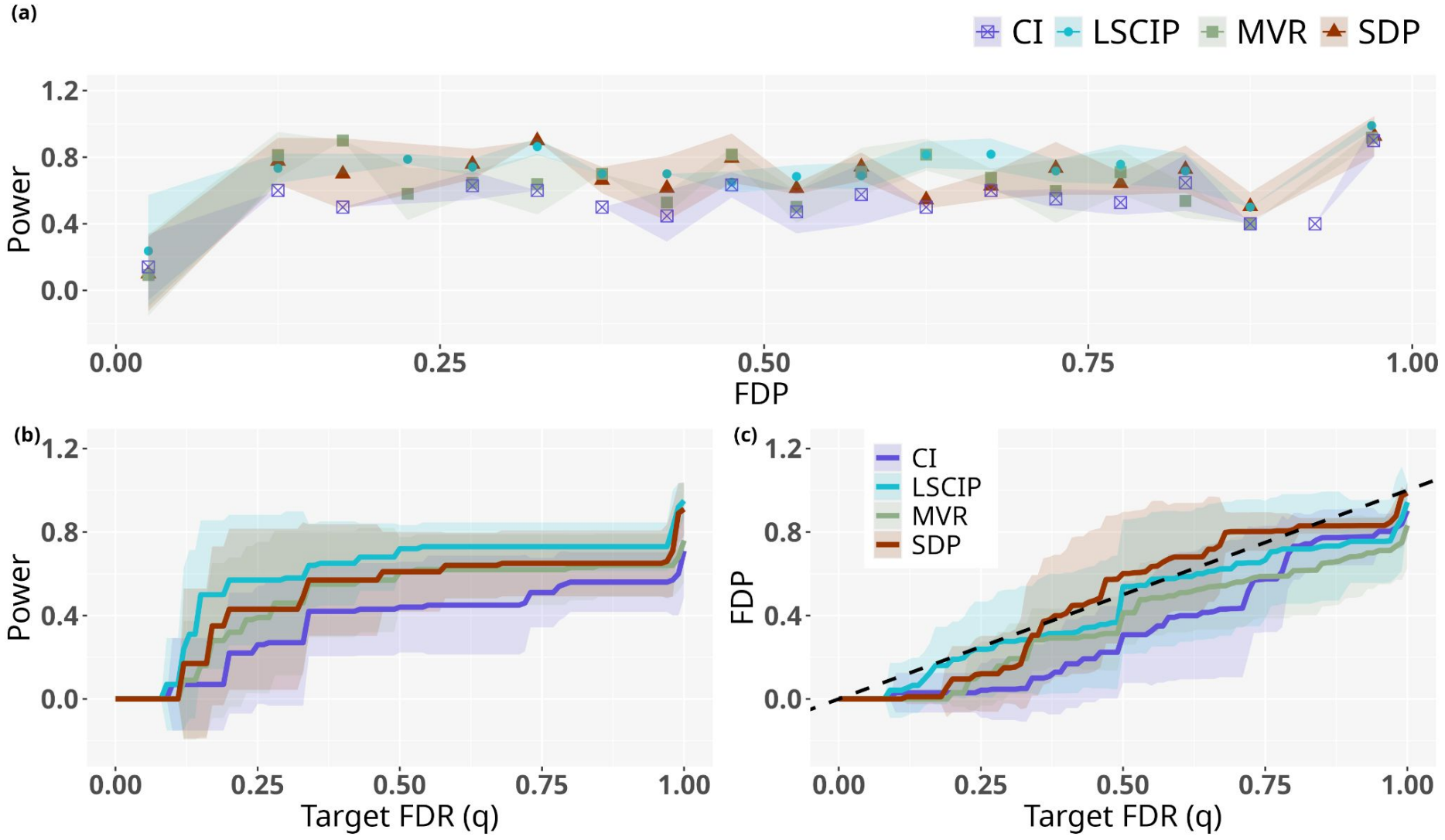
$$\hat{\mathcal{S}} = \{j : W_j \geqslant \tau\}$$

$$FDR := \mathbb{E}[FDP]$$

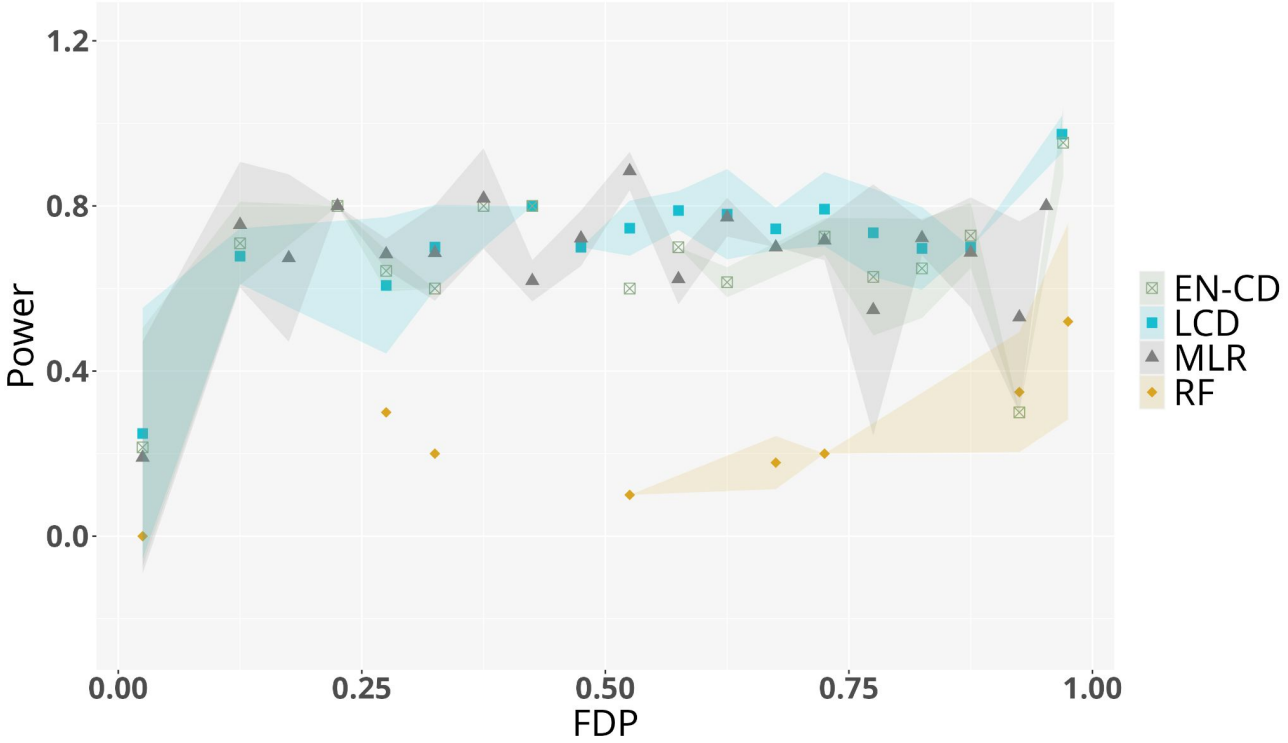
$$FDR = \mathbb{E}\left[\frac{|j \in \hat{\mathcal{S}} \cap \mathcal{H}_0|}{|\hat{\mathcal{S}}| \vee 1}\right] \leqslant q$$

$$\mathcal{H}_0 = \{j \in \{1, .., p\} | Y \perp\!\!\!\perp X_j\}$$

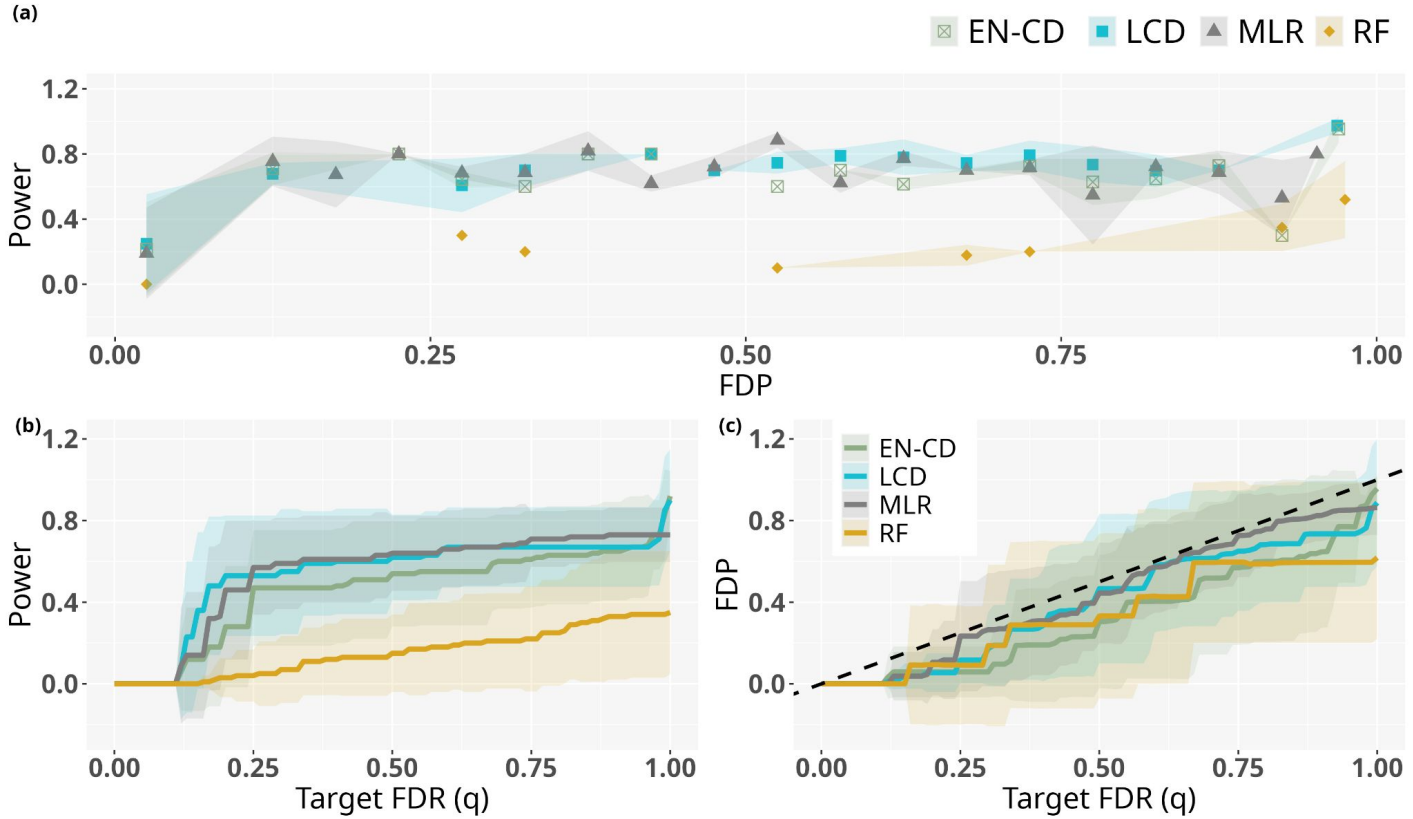
(2) : Candes et al - *Panning for Gold: Model-X Knockoffs for High-dimensional Controlled Variable Selection*

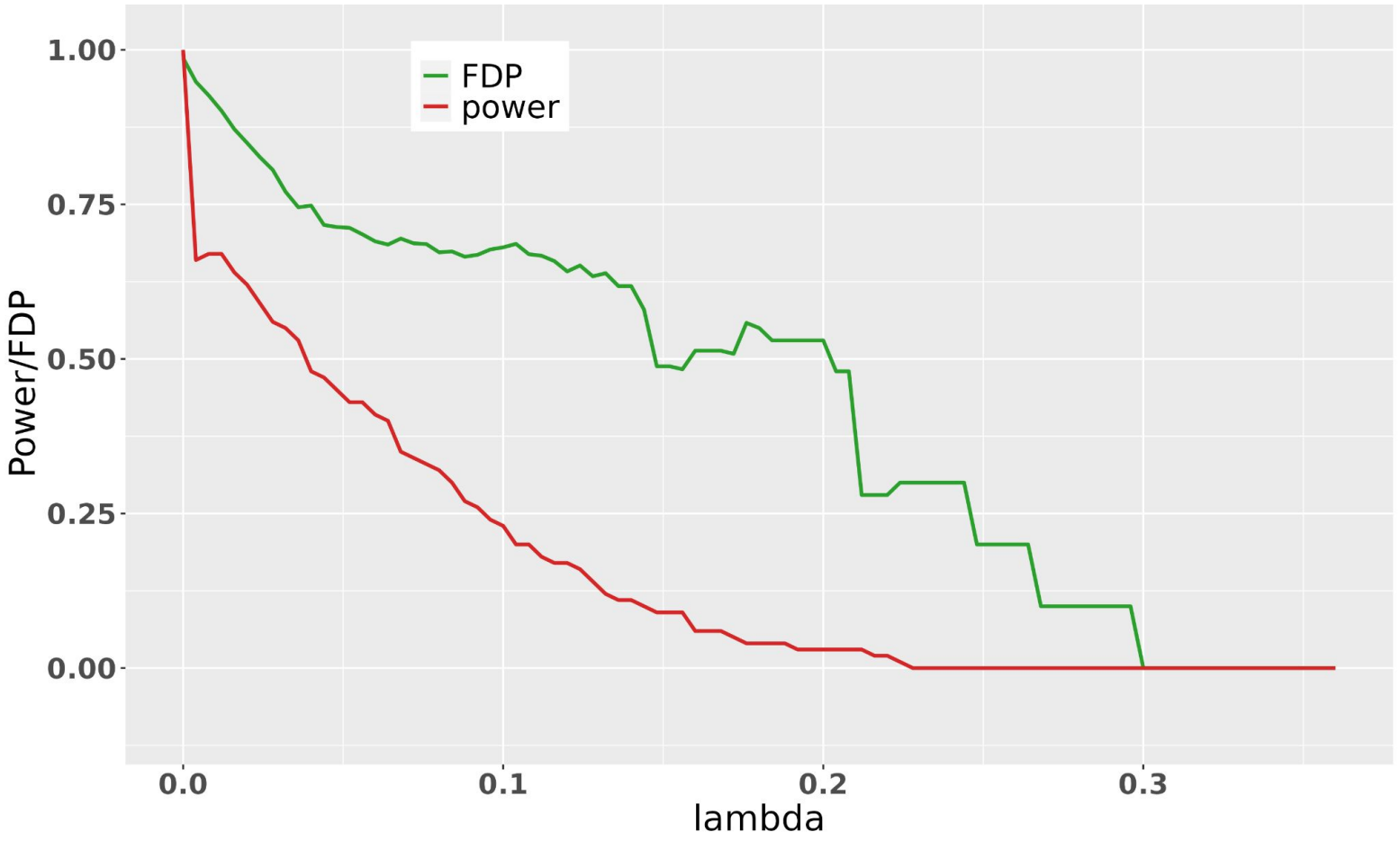

(a)
CI
LSCIP
MVR
SDP
Power
1.2
0.8
0.4
0.0
0.00
0.25
0.50
0.75
1.00
FDP
(b)
Power
Target FDR (q)
(c)
FDP
Target FDR (q)

LSCIP + LCD/MLR based knockoff are the most powerful

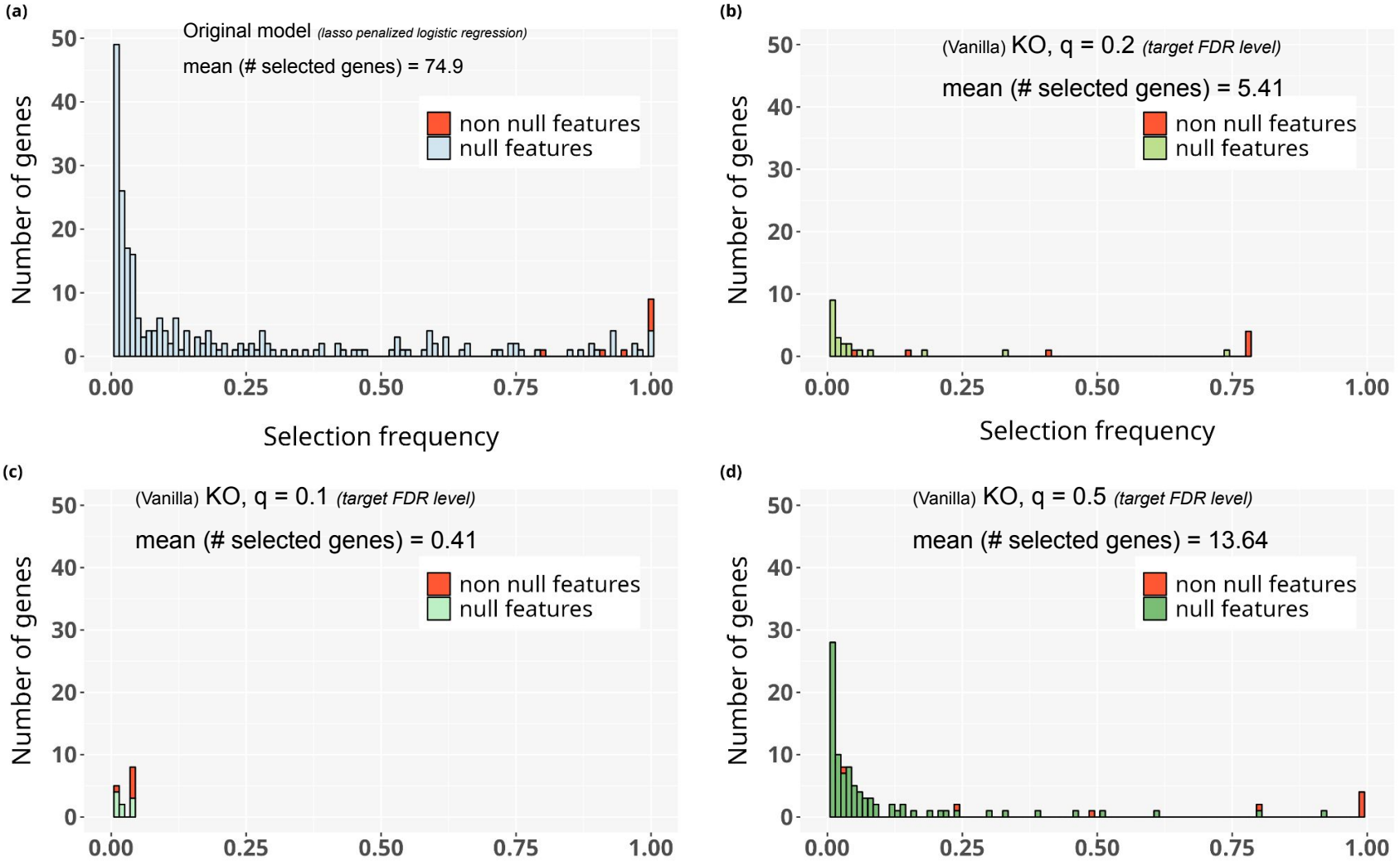


Power and FDP obtained with the knockoff selection procedure

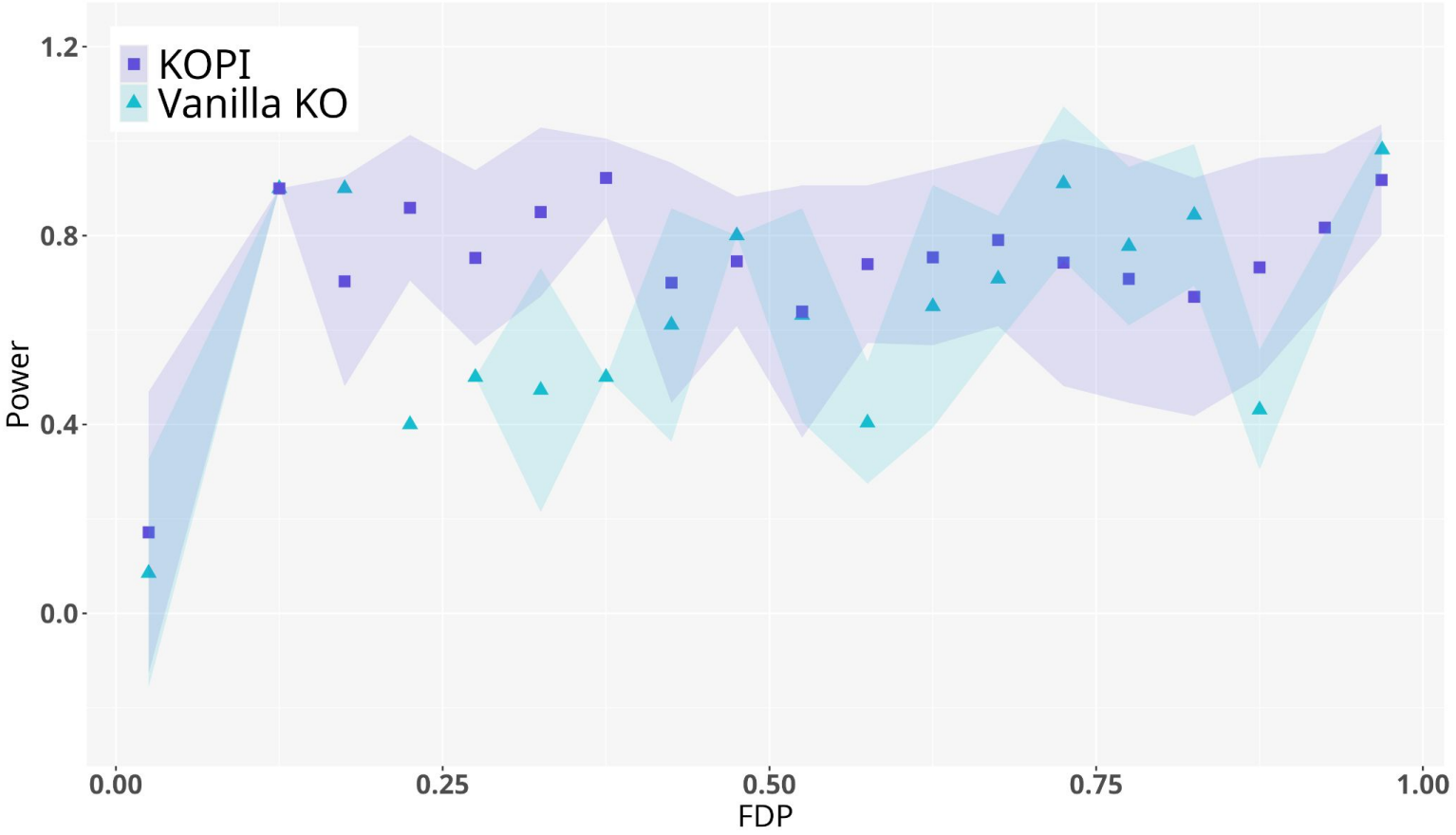

(a)
EN-CD
LCD
MLR
RF
Power
1.2
0.8
0.4
0.0
0.00
0.25
0.50
0.75
1.00
FDP
(b)
Power
Target FDR (q)
(c)
EN-CD
LCD
MLR
RF
FDP
Target FDR (q)


FDP
power
1.00
0.75
0.50
0.25
0.00
Power/FDP
0.0
0.1
0.2
0.3
lambda

Stability selection with simulated data

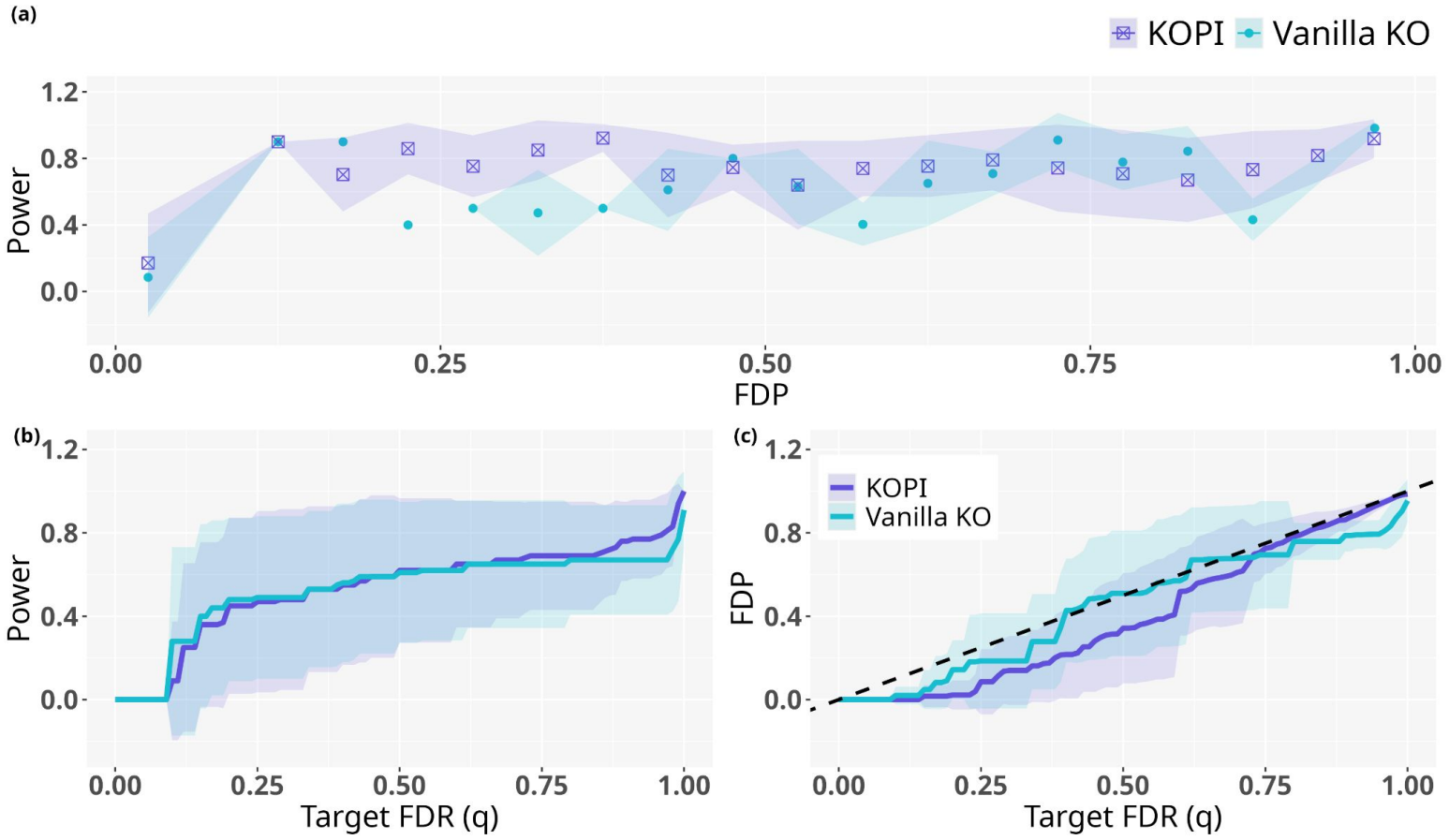


KOPI is empirically at least as powerful as vanilla knockoff

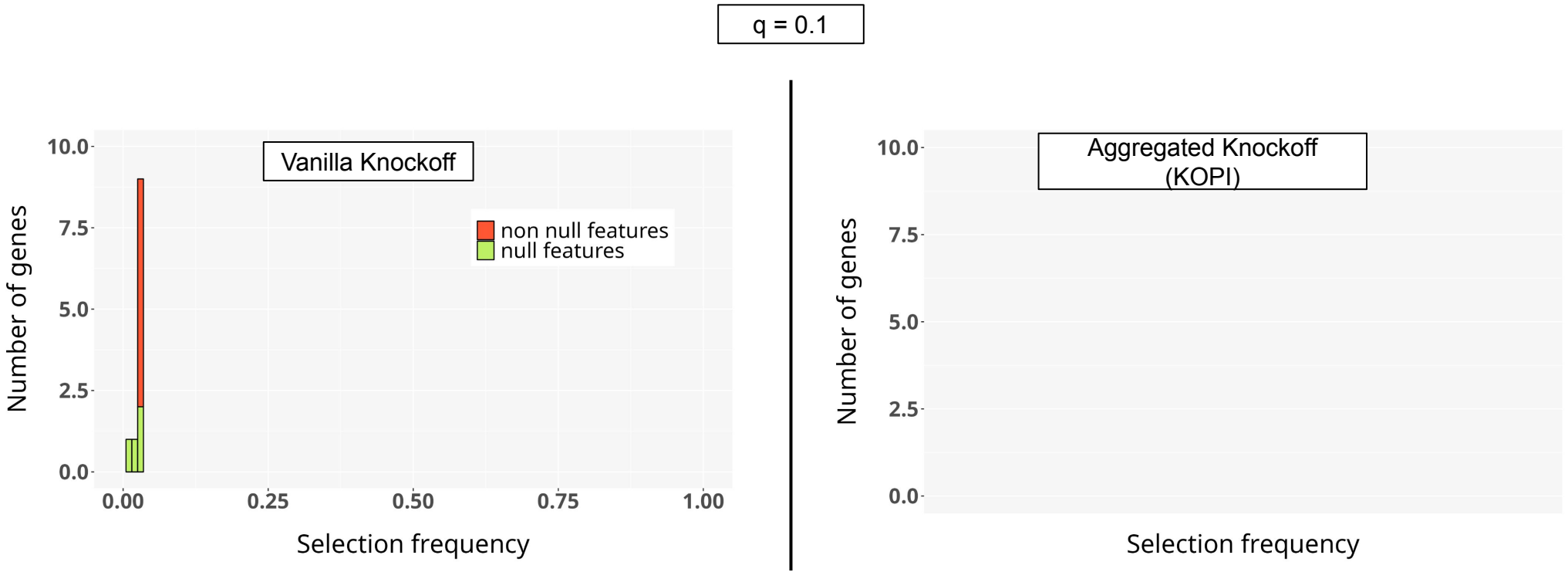


Power and FDP obtained with both Vanilla knockoff and aggregated knockoff selection procedure

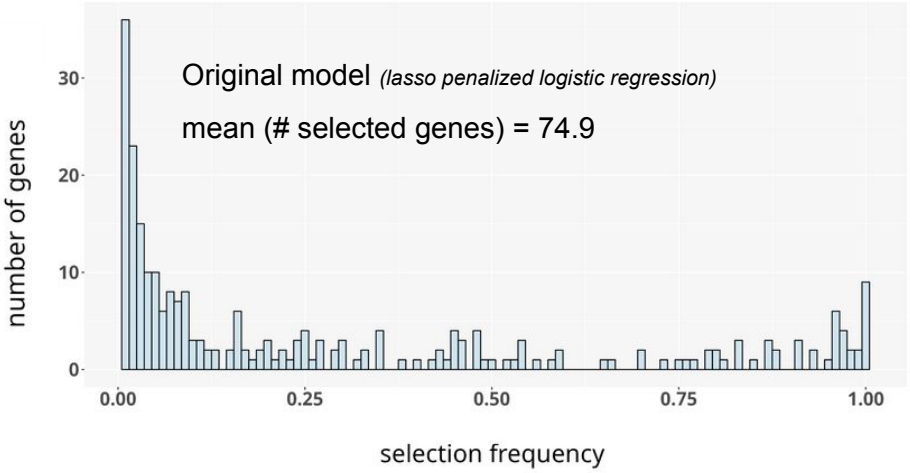
Aggregation simulations

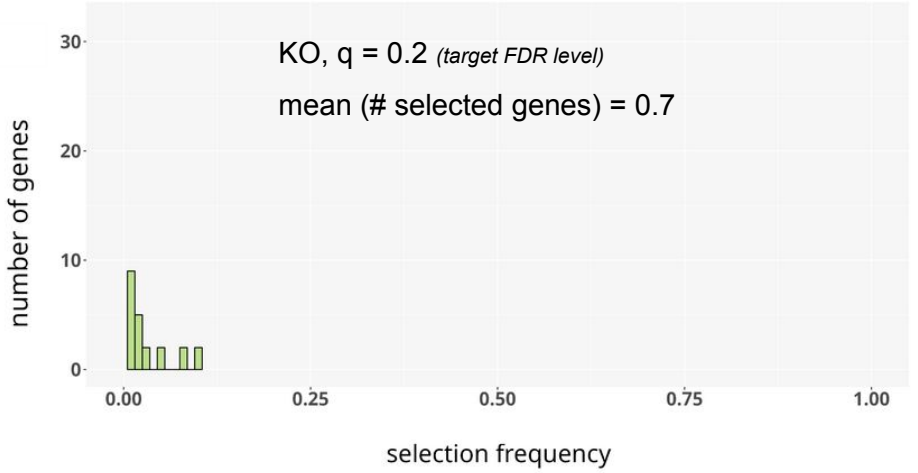


Impact of the aggregation scheme on the selection frequency

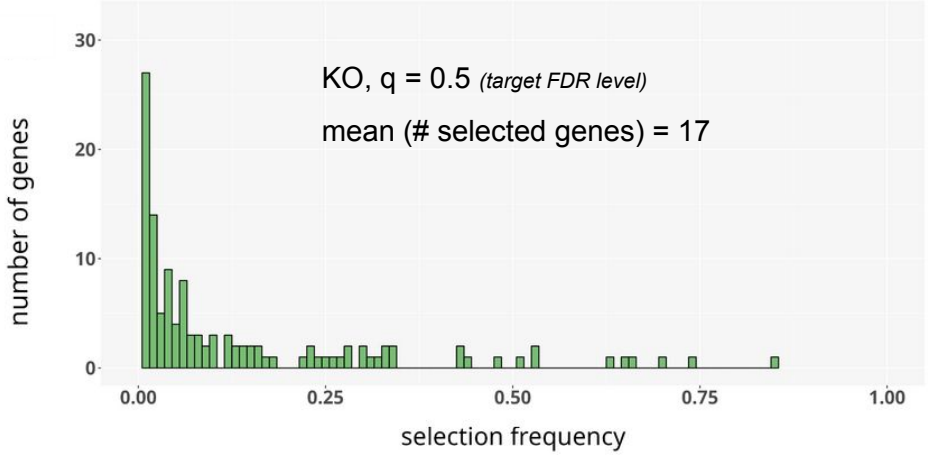


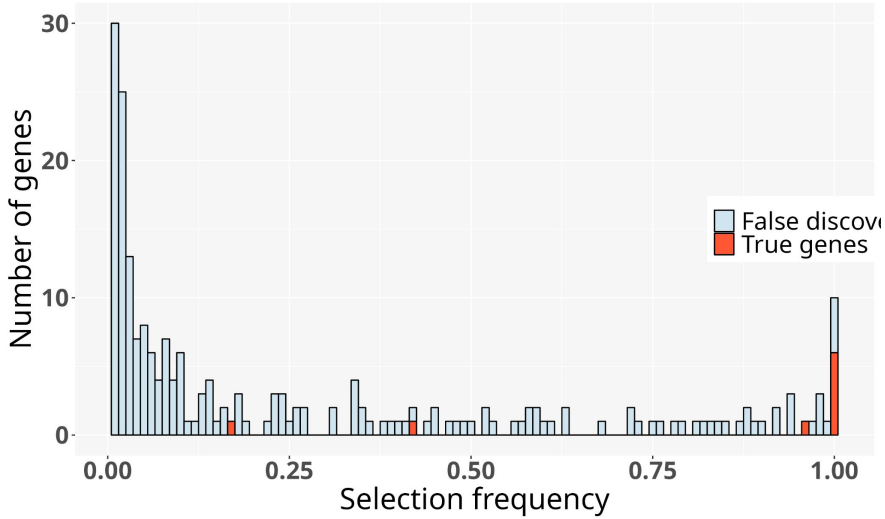
The KO framework does not improve stability (real data with subsampling)

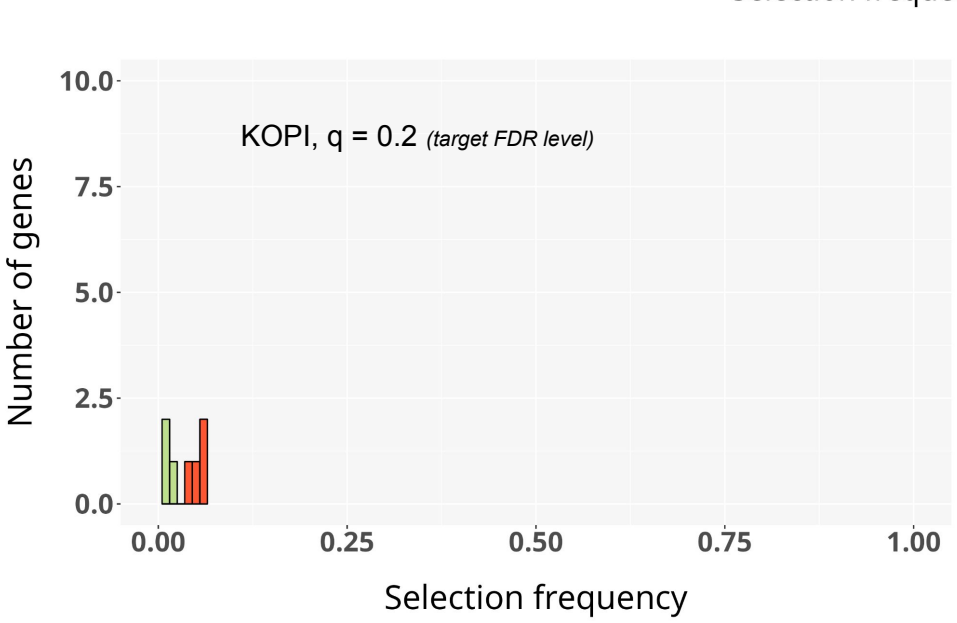


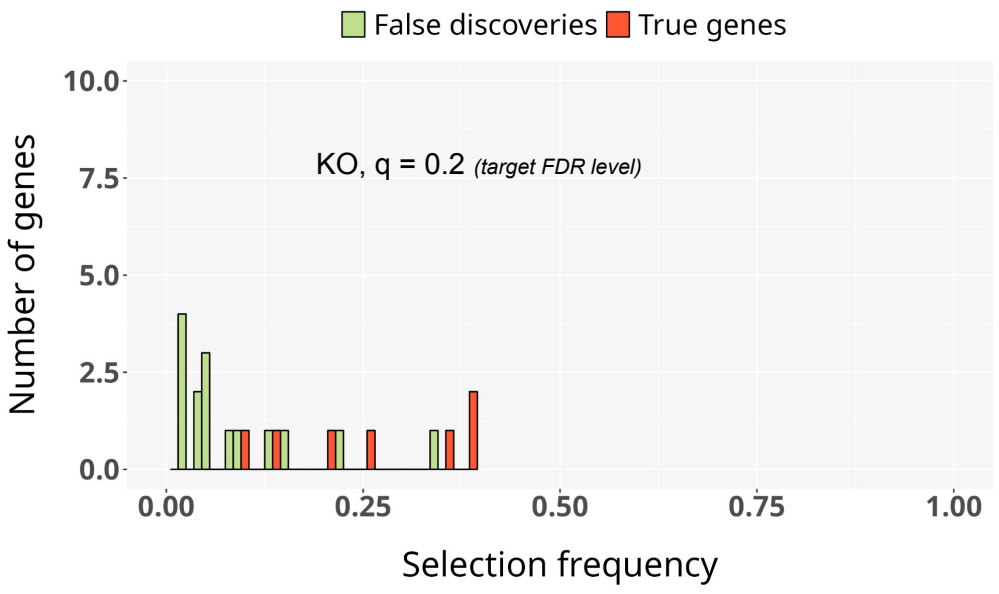


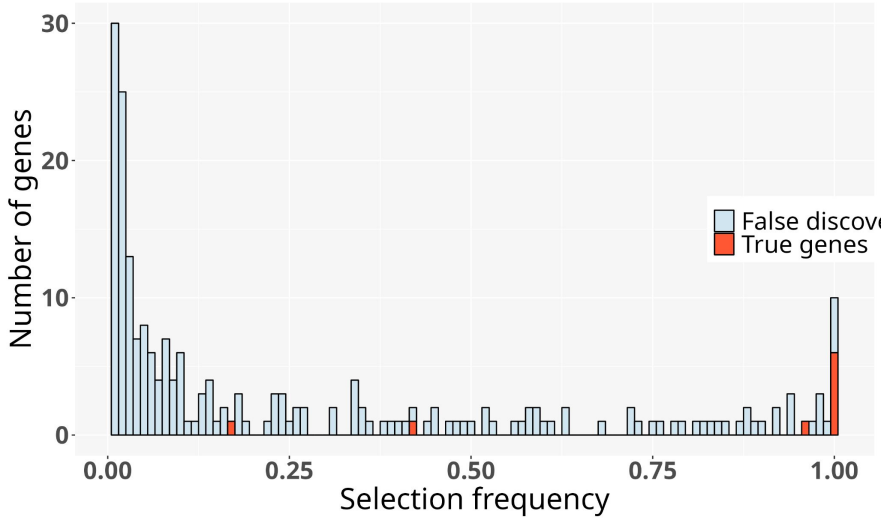
X	real matrix
Y	**cancer/no cancer status**

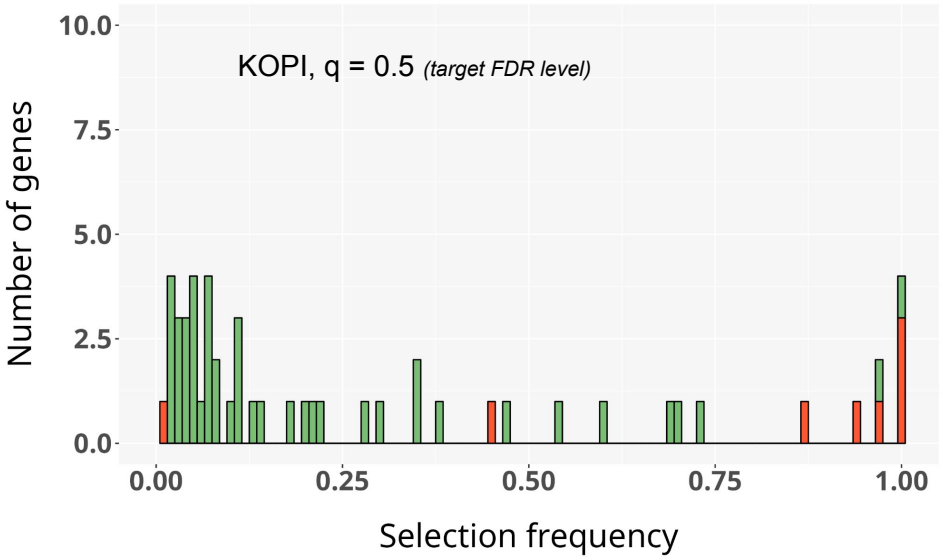


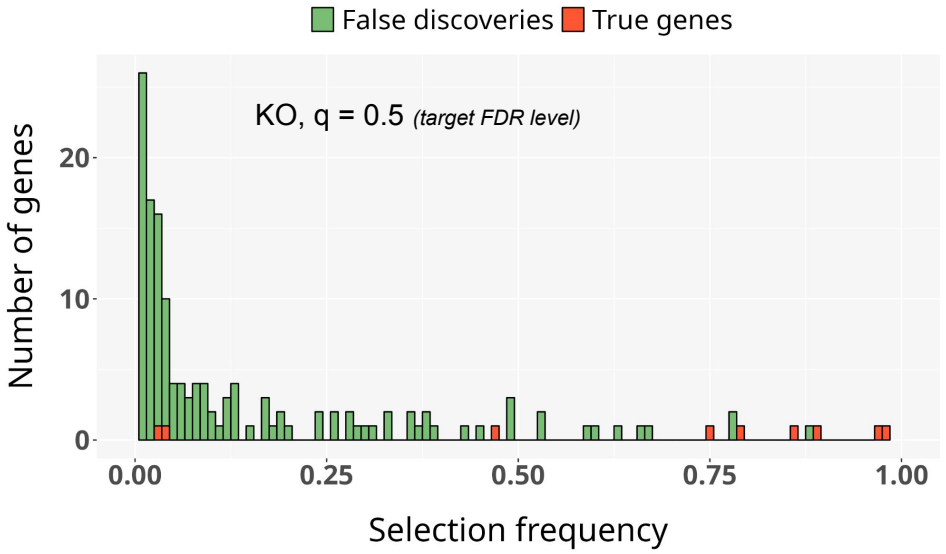

30
20
10
0
Number of genes
0.00
0.25
0.50
0.75
1.00
Selection frequency
False discov
True genes


10.0
7.5
5.0
2.5
0.0
Number of genes
KOPI, q = 0.2 (target FDR level)
0.00
0.25
0.50
0.75
1.00
Selection frequency

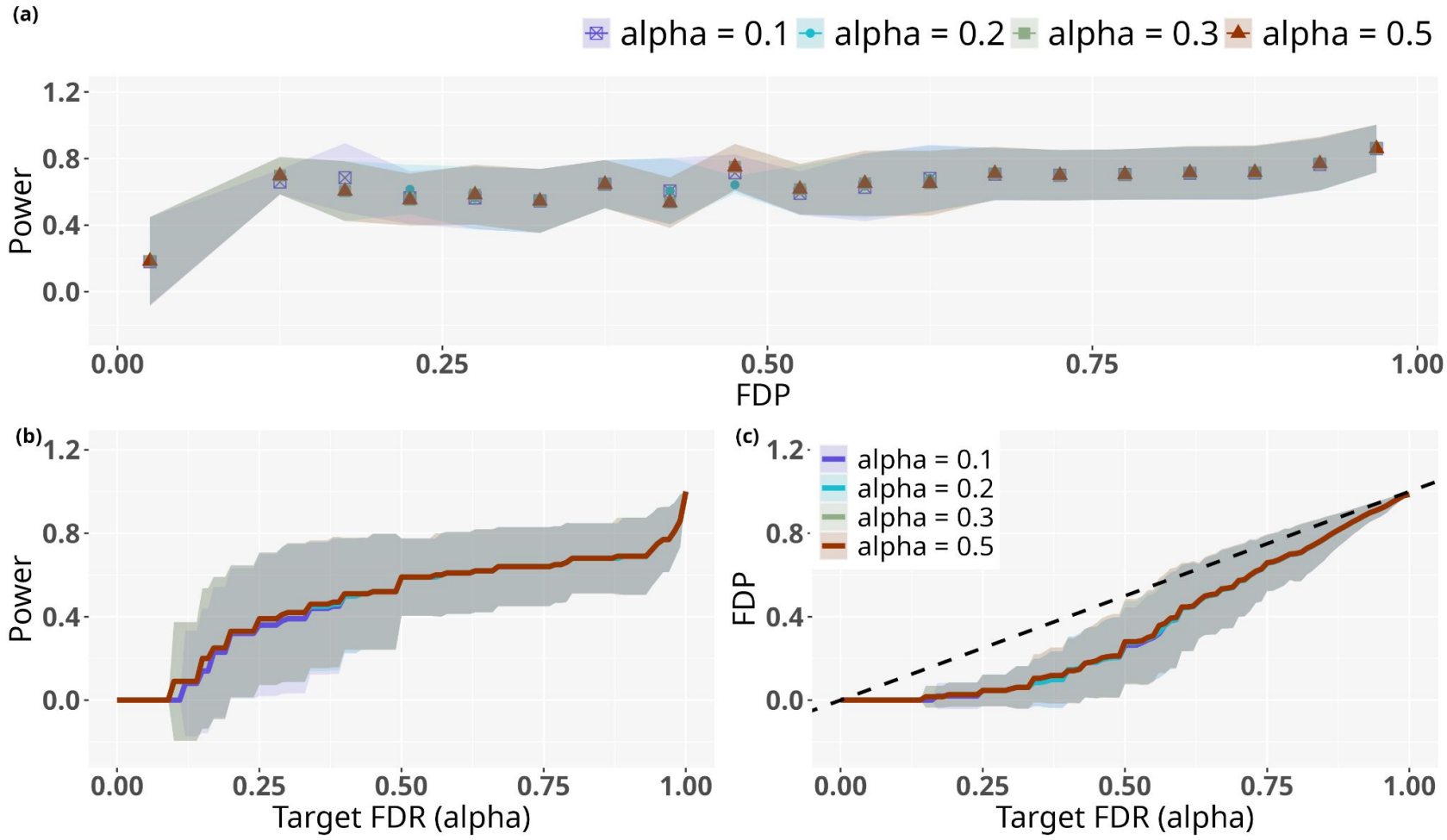

False discoveries
True genes
10.0
7.5
5.0
2.5
0.0
Number of genes
KO, q = 0.2 (target FDR level)
0.00
0.25
0.50
0.75
1.00
Selection frequency


30
20
10
0
Number of genes
False discov
True genes
0.00
0.25
0.50
0.75
1.00
Selection frequency


KOPI, q = 0.5 (target FDR level)
10.0
7.5
5.0
2.5
0.0
Number of genes
0.00
0.25
0.50
0.75
1.00
Selection frequency


False discoveries
True genes
KO, q = 0.5 (target FDR level)
20
10
0
Number of genes
0.00
0.25
0.50
0.75
1.00
Selection frequency

Alpha effect



Study the applicability of the knockoff procedure to transcriptomic data in classification

→ **Identified limitations of the knockoff framework**

The knockoff procedure fails with interacting variables

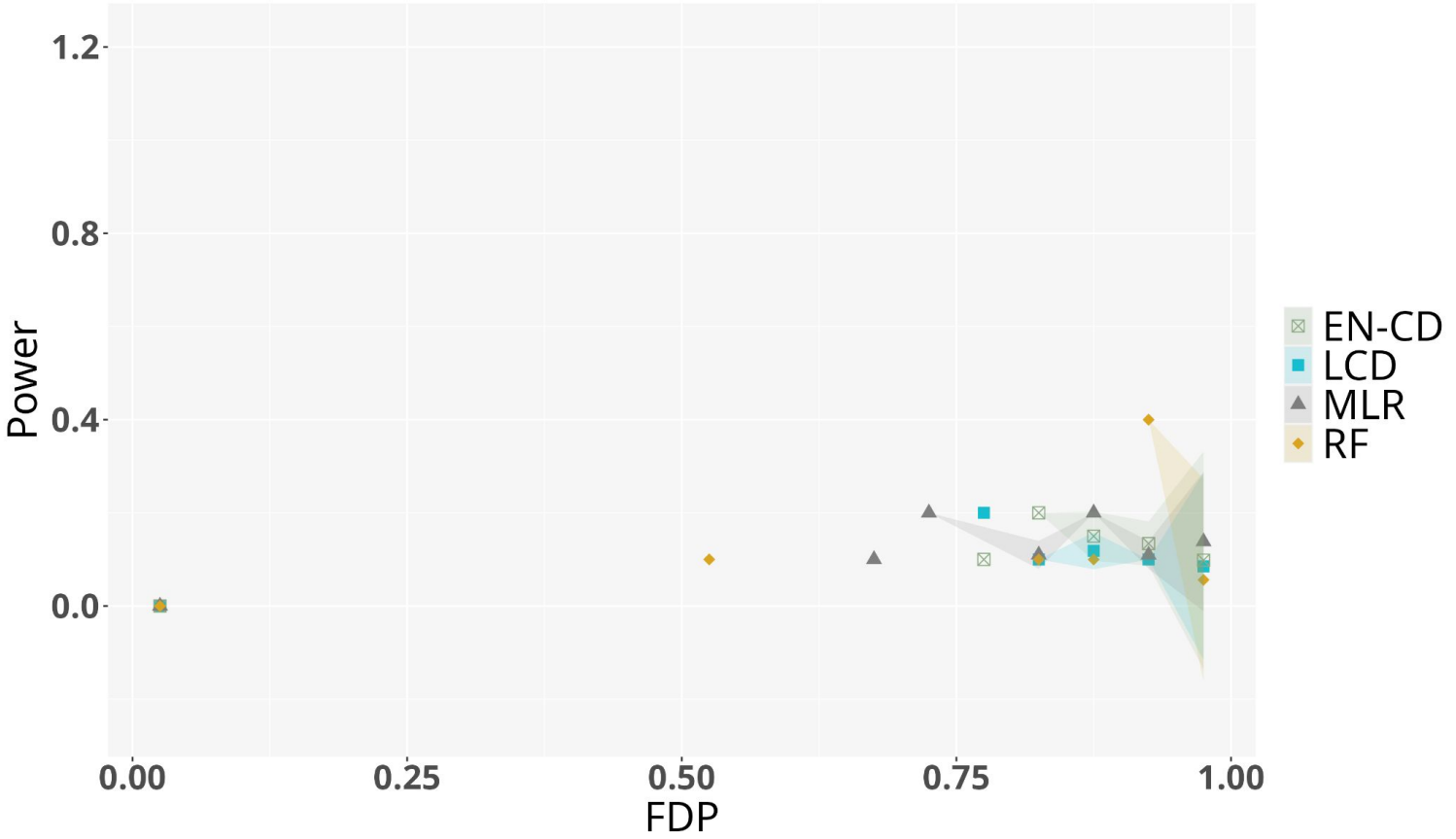
X	**Real matrix**
Y	**Interaction**

$$Y = \mathbb{1}_{\left(\frac{1}{1+e^{-X_n\beta}} > 0.5\right)}$$

$$X_n\beta = \beta_1 X_1 X_2 + \beta_2 X_3 X_4 + \cdots + \beta_5 X_9 X_{10}$$

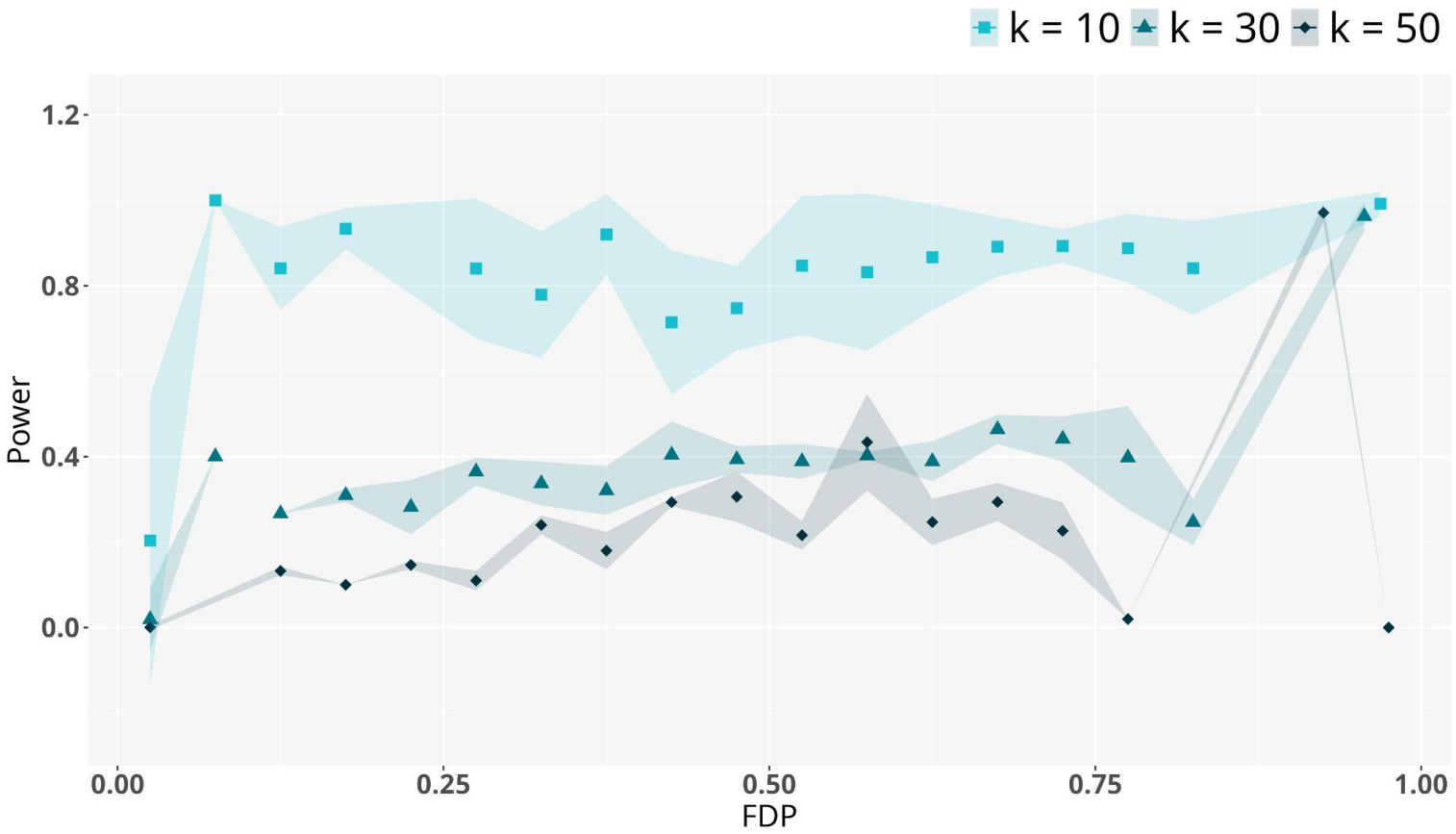
The knockoff procedure fails with interacting variables

X	Real matrix
Y	**Interaction**



Power obtained with the knockoff selection procedure

The power decreases with the number of non null features

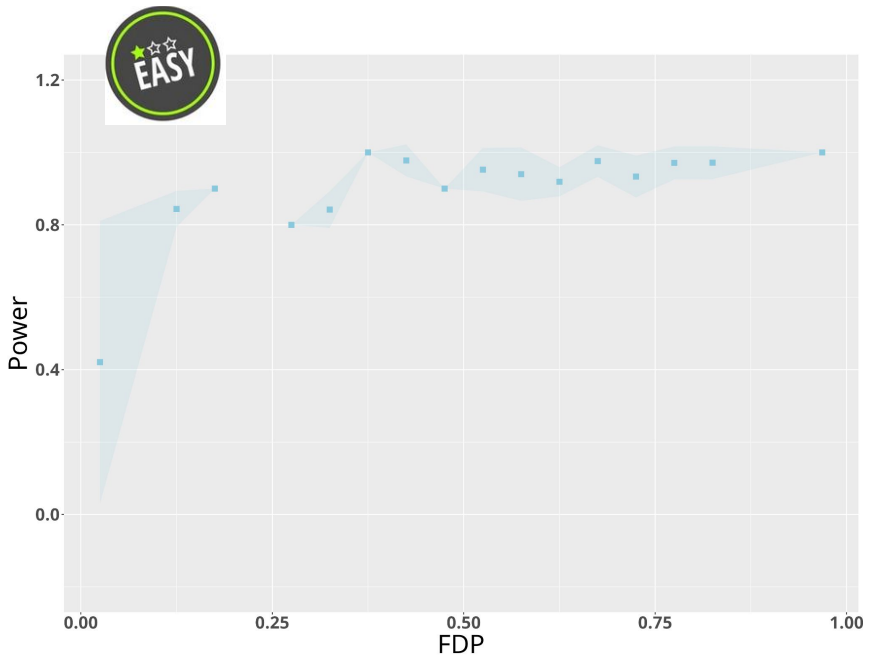


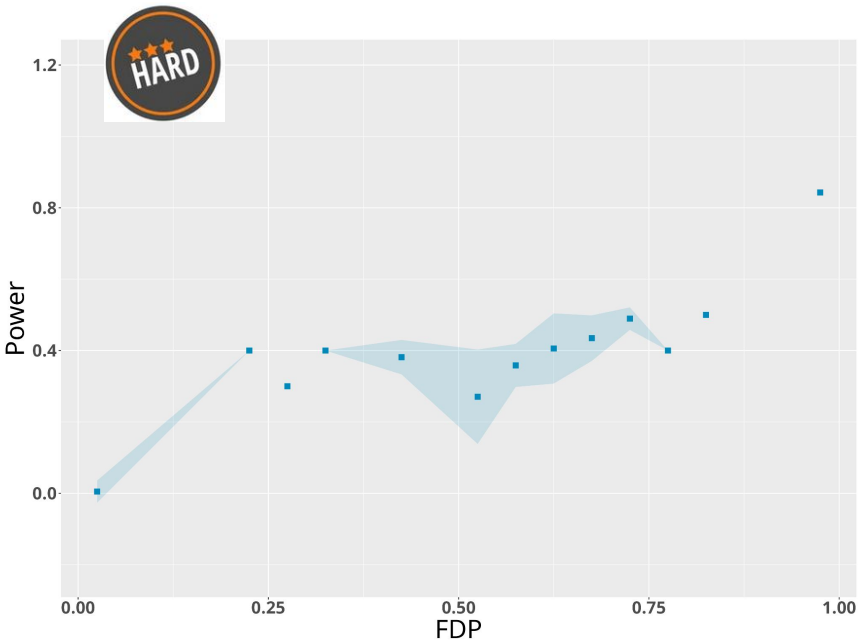
Power and FDP obtained with the LSCIP + LCD knockoff selection procedure **for different numbers of non-null features (k)**

Performances also depend on the features used to simulate Y…

- Setting (linear)
- KO framework (LSCIP + LCD)
- Features matrix

- **Non null features (10)**

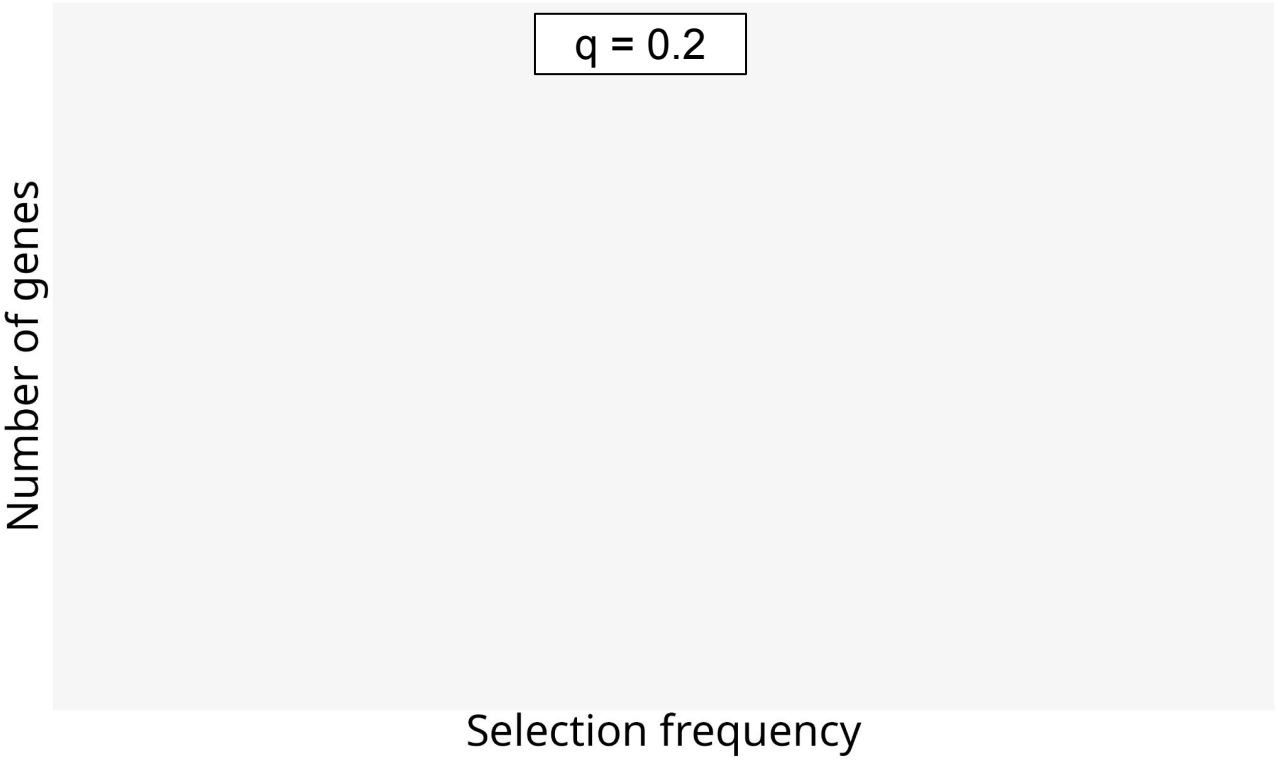




Power and FDP obtained with the LSCIP + LCD knockoff selection procedure

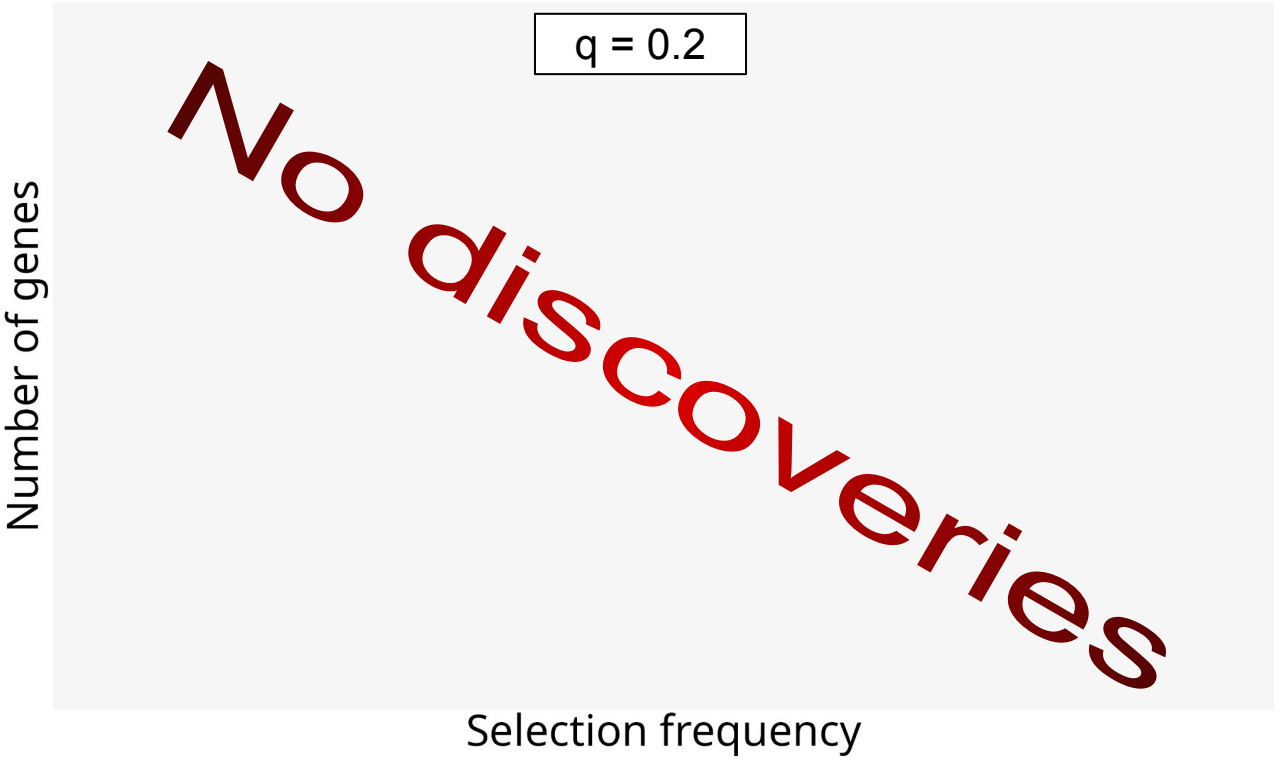
Application to the real data

X	Real matrix
Y	**Cancer/no cancer**



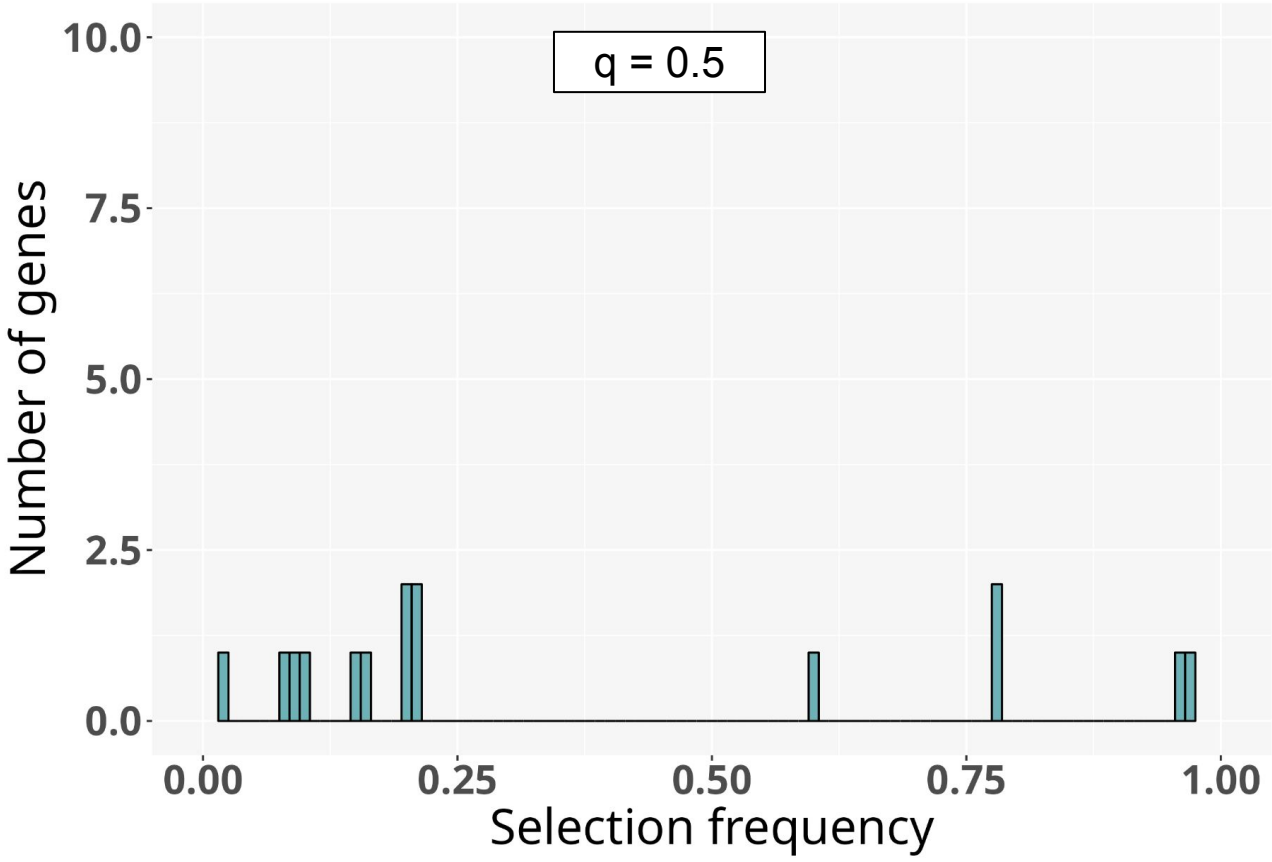
Application to the real data

X	Real matrix
Y	**Cancer/no cancer**

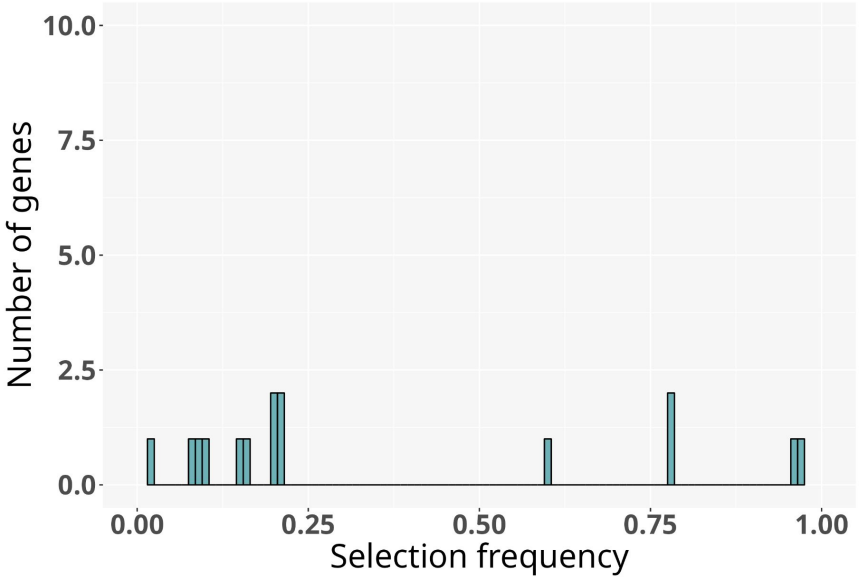


Application to the real data

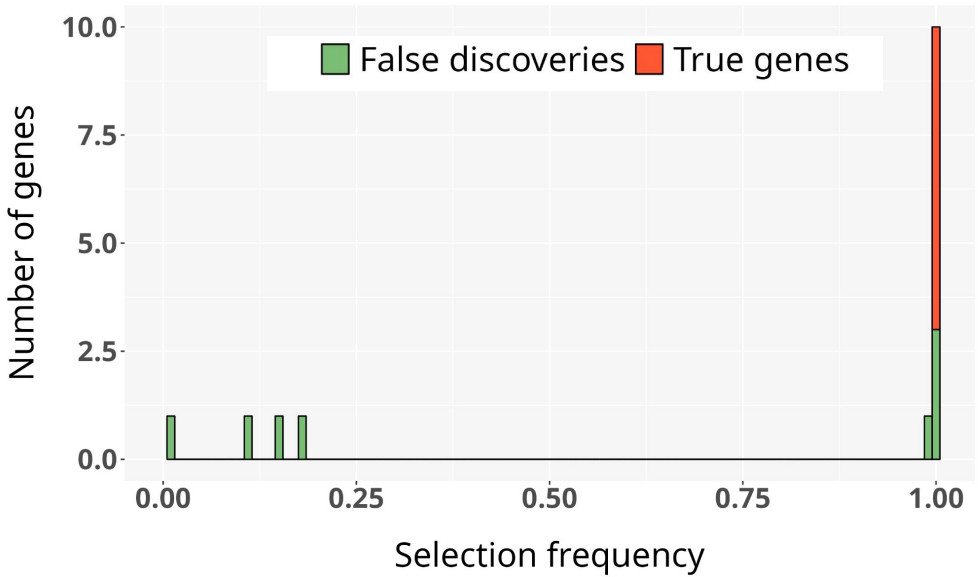
X	Real matrix
Y	**Cancer/no cancer**

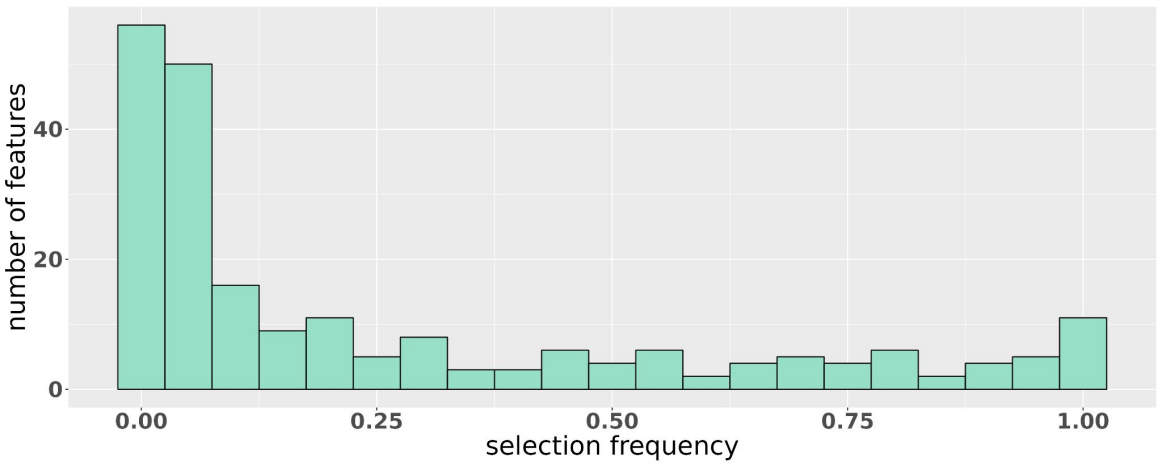


Application to the real data vs simulated data



q = 0.5





(3) : de Biase et al. - *Smoking-dependent expression alterations in the general population reveal immune impairment linked to germline variation and lung cancer risk*