The Multilayer Knockoff Filter: Controlled Multi-Resolution Variable Selection

Eugene Katsevich

June 21, 2017

Section 1

Model selection at multiple resolutions

Genetic association studies

Data:

- Phenotype measurements $\boldsymbol{y} \in \mathbb{R}^n$.
- Genotype measurements $\boldsymbol{X} \in \mathbb{R}^{n \times p}$.

Scientific question:

- Which single nucleotide polymorphisms (SNPs) are associated to the phenotype?

A typical GWAS output table

Table 2 SNPs showing association with Alzheimer's disease at $P \leq 1 \times 10^{-5}$

SNP	Chr	MB	Closest RefSeq gene	Location relative to gene	GWAS *P* value
rs11894266	2	170.3	*SSB*	5′	6.9×10^{-7}
rs610932	11	59.7	*MS4A6A*	3′ UTR	1.4×10^{-6}
rs10501927	11	99.3	*CNTN5*	Intronic	2.0×10^{-6}
rs9446432	6	72.4	*C6orf155*	Intergenic	2.8×10^{-6}
rs7561528	2	127.6	*BIN1*	5′	3.0×10^{-6}
rs744373	2	127.6	*BIN1*	5′	3.2×10^{-6}
rs662196	11	59.7	*MS4A6A*	Intronic	5.2×10^{-6}
rs583791	11	59.7	*MS4A6A*	Intronic	5.3×10^{-6}
rs676309	11	59.8	*MS4A4E*	5′	6.3×10^{-6}
rs1157242	8	37.2	*KCNU1*	Intergenic	7.0×10^{-6}
rs1539053	1	57.9	*DAB1*	Intronic	7.1×10^{-6}
rs11827375	11	76.0	*C11orf30*	3′	7.2×10^{-6}
rs1408077	1	205.9	*CR1*	Intronic	8.3×10^{-6}
rs9384428	6	156.5	*ARID1B*	Intergenic	8.5×10^{-6}
rs6701713	1	205.9	*CR1*	Intronic	8.7×10^{-6}
rs3818361	1	205.9	*CR1*	Intronic	9.2×10^{-6}

Figure: Source: Harold et al. Nature Genetics 41.10 (2009): 1088.

A typical GWAS output table

Table 2 SNPs showing association with Alzheimer's disease at $P \leq 1 \times 10^{-5}$

SNP	Chr	MB	Closest RefSeq gene	Location relative to gene	GWAS *P* value
rs11894266	2	170.3	*SSB*	5′	6.9×10^{-7}
rs610932	11	59.7	*MS4A6A*	3′ UTR	1.4×10^{-6}
rs10501927	11	99.3	*CNTN5*	Intronic	2.0×10^{-6}
rs9446432	6	72.4	*C6orf155*	Intergenic	2.8×10^{-6}
rs7561528	2	127.6	*BIN1*	5′	3.0×10^{-6}
rs744373	2	127.6	*BIN1*	5′	3.2×10^{-6}
rs662196	11	59.7	*MS4A6A*	Intronic	5.2×10^{-6}
rs583791	11	59.7	*MS4A6A*	Intronic	5.3×10^{-6}
rs676309	11	59.8	*MS4A4E*	5′	6.3×10^{-6}
rs1157242	8	37.2	*KCNU1*	Intergenic	7.0×10^{-6}
rs1539053	1	57.9	*DAB1*	Intronic	7.1×10^{-6}
rs11827375	11	76.0	*C11orf30*	3′	7.2×10^{-6}
rs1408077	1	205.9	*CR1*	Intronic	8.3×10^{-6}
rs9384428	6	156.5	*ARID1B*	Intergenic	8.5×10^{-6}
rs6701713	1	205.9	*CR1*	Intronic	8.7×10^{-6}
rs3818361	1	205.9	*CR1*	Intronic	9.2×10^{-6}

Figure: Source: Harold et al. Nature Genetics 41.10 (2009): 1088.

We have discoveries at both the SNP and gene levels.

A typical GWAS output table

Table 2 SNPs showing association with Alzheimer's disease at $P \leq 1 \times 10^{-5}$

SNP	Chr	MB	Closest RefSeq gene	Location relative to gene	GWAS *P* value
rs11894266	2	170.3	*SSB*	5′	6.9×10^{-7}
rs610932	11	59.7	*MS4A6A*	3′ UTR	1.4×10^{-6}
rs10501927	11	99.3	*CNTN5*	Intronic	2.0×10^{-6}
rs9446432	6	72.4	*C6orf155*	Intergenic	2.8×10^{-6}
rs7561528	2	127.6	*BIN1*	5′	3.0×10^{-6}
rs744373	2	127.6	*BIN1*	5′	3.2×10^{-6}
rs662196	11	59.7	*MS4A6A*	Intronic	5.2×10^{-6}
rs583791	11	59.7	*MS4A6A*	Intronic	5.3×10^{-6}
rs676309	11	59.8	*MS4A4E*	5′	6.3×10^{-6}
rs1157242	8	37.2	*KCNU1*	Intergenic	7.0×10^{-6}
rs1539053	1	57.9	*DAB1*	Intronic	7.1×10^{-6}
rs11827375	11	76.0	*C11orf30*	3′	7.2×10^{-6}
rs1408077	1	205.9	*CR1*	Intronic	8.3×10^{-6}
rs9384428	6	156.5	*ARID1B*	Intergenic	8.5×10^{-6}
rs6701713	1	205.9	*CR1*	Intronic	8.7×10^{-6}
rs3818361	1	205.9	*CR1*	Intronic	9.2×10^{-6}

Figure: Source: Harold et al. Nature Genetics 41.10 (2009): 1088.

We have discoveries at both the SNP and gene levels. $\Rightarrow$
We seek to control both corresponding FDRs.

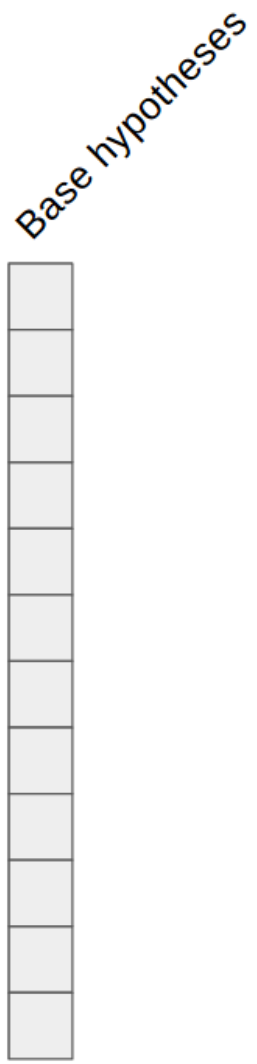
Model selection at multiple layers

- Outcome variable y and predictors $X_1, \ldots, X_p$.

Model selection at multiple layers

- Outcome variable y and predictors $X_1, \ldots, X_p$.
- Base-level hypotheses $H_1, \ldots, H_p$, where

$$H_j : y \perp\!\!\!\perp X_j | X_{-j}.$$



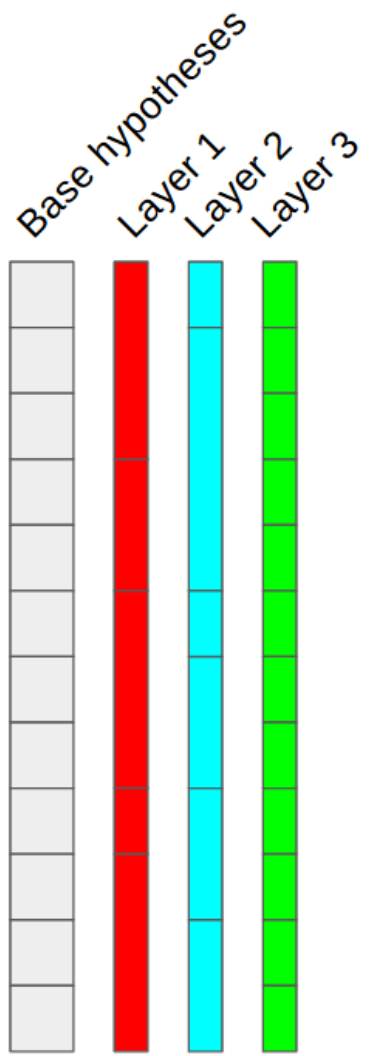
Model selection at multiple layers

- Outcome variable y and predictors $X_1, \ldots, X_p$.
- Base-level hypotheses $H_1, \ldots, H_p$, where

$$H_j : y \perp\!\!\!\perp X_j | X_{-j}.$$

- For each $m = 1, \ldots, M$, partition hypotheses into disjoint groups $\mathcal{A}_g^m$:

$$\{1, \ldots, p\} = \bigcup_{g=1}^{G_m} \mathcal{A}_g^m.$$



Model selection at multiple layers

- Outcome variable y and predictors $X_1, \ldots, X_p$.
- Base-level hypotheses $H_1, \ldots, H_p$, where

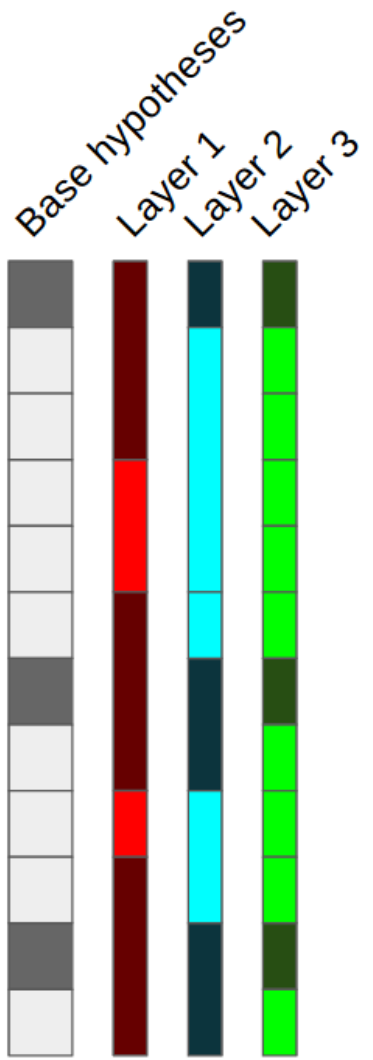
$$H_j : y \perp\!\!\!\perp X_j | X_{-j}.$$

- For each $m = 1, \ldots, M$, partition hypotheses into disjoint groups $\mathcal{A}_g^m$:

$$\{1, \ldots, p\} = \bigcup_{g=1}^{G_m} \mathcal{A}_g^m.$$

- Selection set $\mathcal{S}$ induces selections at each layer:

$$\mathcal{S}_m = \{g = 1, \ldots, G_m : A_g^m \text{ intersects } \mathcal{S}\}.$$



Multilayer FDR control[1]

Definition

A model selection procedure obeys **multilayer FDR control** at target levels $q_1, \ldots, q_M$ if

$$\mathrm{FDR}_m = \mathbb{E}\left[\frac{|\mathcal{S}_m \cap \mathcal{H}_0^m|}{|\mathcal{S}_m|}\right] \leq q_m \quad \text{for all } m.$$

[1] Barber and Ramdas '15

Section 2

Building blocks: p-filter and knockoff filter

p-filter[2]

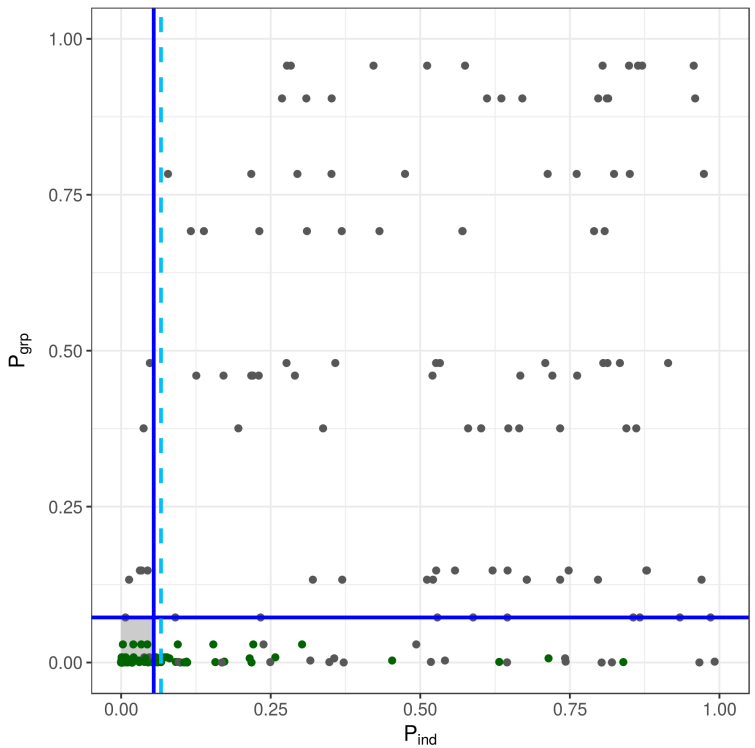
If p-values for base-level hypotheses are available...

[2] Barber and Ramdas '15

p-filter[2]

If p-values for base-level hypotheses are available...

1. Get group p-values using Simes.

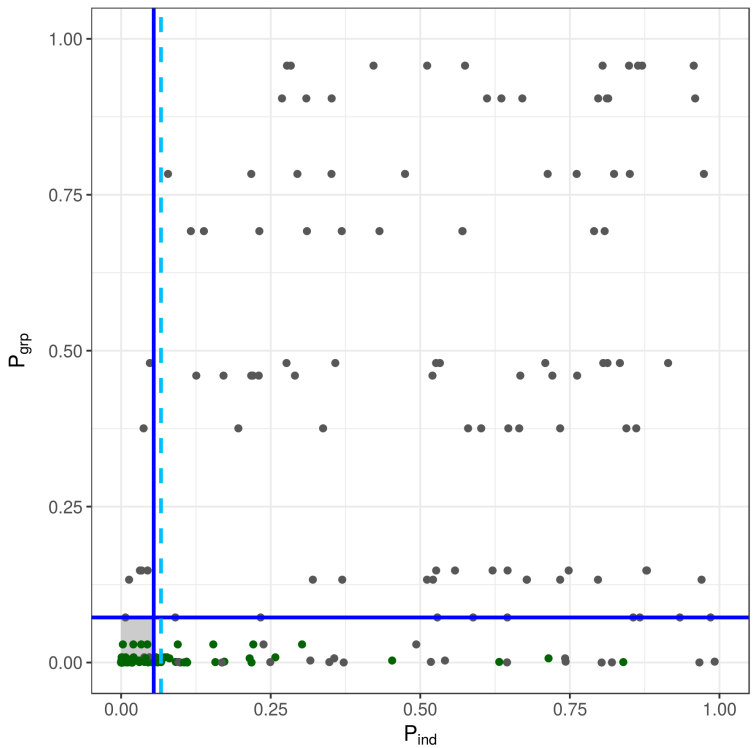


[2] Barber and Ramdas '15

p-filter[2]

If p-values for base-level hypotheses are available...

1. Get group p-values using Simes.
2. Introduce thresholds $\boldsymbol{t} = (t_1, \ldots, t_M)$.

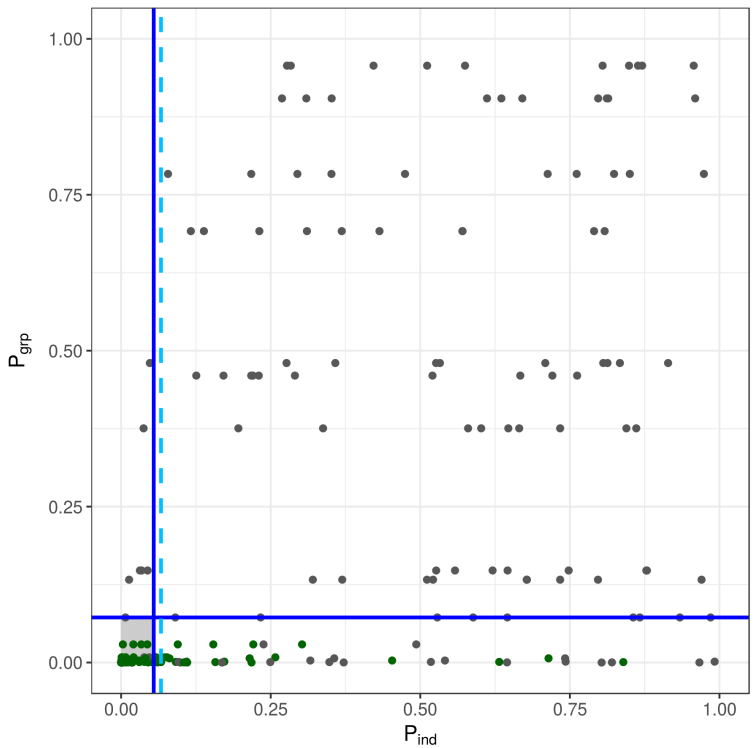


[2] Barber and Ramdas '15

p-filter[2]

If p-values for base-level hypotheses are available...

1. Get group p-values using Simes.
2. Introduce thresholds $\boldsymbol{t} = (t_1, \ldots, t_M)$.
3. Select hypotheses $\mathcal{S}(\boldsymbol{t})$ passing thresholds at all layers.



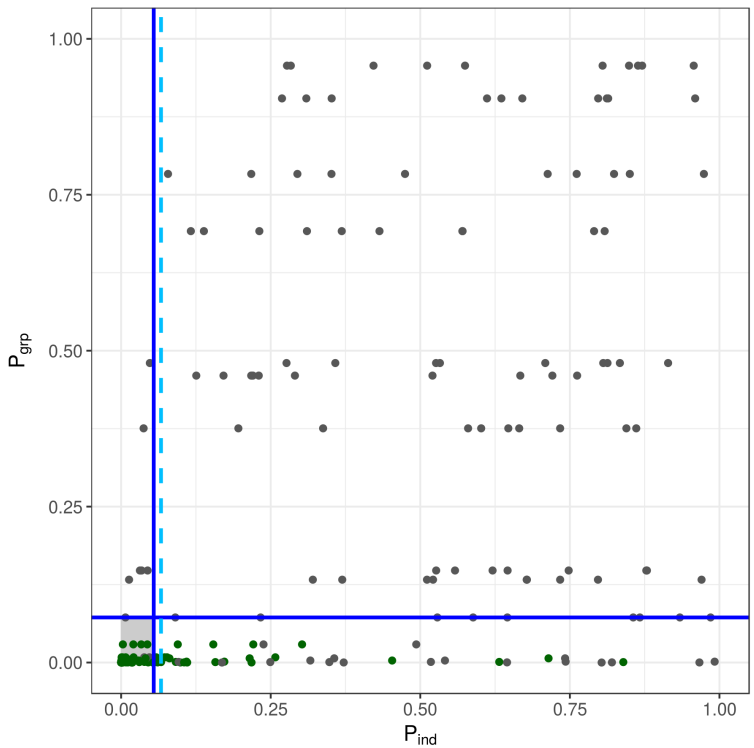
[2] Barber and Ramdas '15

p-filter[2]

If p-values for base-level hypotheses are available...

1. Get group p-values using Simes.
2. Introduce thresholds $\boldsymbol{t} = (t_1, \ldots, t_M)$.
3. Select hypotheses $\mathcal{S}(\boldsymbol{t})$ passing thresholds at all layers.
4. Choose

$$\boldsymbol{t}^* = \max\{\boldsymbol{t} : \widehat{\mathrm{FDP}}_m(\boldsymbol{t}) \leq q_m \; \forall m\}.$$



[2] Barber and Ramdas '15

Knockoff filter[3]

A model selection procedure bypassing the construction of p-values.

[3] Barber and Candes '15

Knockoff filter[3]

A model selection procedure bypassing the construction of p-values.

1. Construct "knockoff variables" $\widetilde{\boldsymbol{X}}$ to use as controls.

[3] Barber and Candes '15

Knockoff filter[3]

A model selection procedure bypassing the construction of p-values.

1. Construct "knockoff variables" $\widetilde{\boldsymbol{X}}$ to use as controls.
2. Create statistics $\boldsymbol{W} = (W_1, \ldots, W_p)$, where W_j quantifies how much more "significant" X_j is than $\widetilde{X}_j$.

[3] Barber and Candes '15

Knockoff filter[3]

A model selection procedure bypassing the construction of p-values.

1. Construct "knockoff variables" $\widetilde{\boldsymbol{X}}$ to use as controls.
2. Create statistics $\boldsymbol{W} = (W_1, \ldots, W_p)$, where W_j quantifies how much more "significant" X_j is than $\widetilde{X}_j$.
3. Consider $\mathcal{S}(t) = \{j : W_j \geq t\}$.

[3] Barber and Candes '15

Knockoff filter[3]

A model selection procedure bypassing the construction of p-values.

1. Construct "knockoff variables" $\widetilde{\boldsymbol{X}}$ to use as controls.
2. Create statistics $\boldsymbol{W} = (W_1, \ldots, W_p)$, where W_j quantifies how much more "significant" X_j is than $\widetilde{X}_j$.
3. Consider $\mathcal{S}(t) = \{j : W_j \geq t\}$.
4. Select $t = \min\{t : \widehat{\mathrm{FDP}}(t) \leq q\}$.

[3] Barber and Candes '15

Section 3

Multilayer knockoff filter

A synthesis of the two approaches

I propose the **multilayer knockoff filter (MKF)**, which leverages

- The multilayer testing framework of the p-filter;
- Test statistics for model selection from the knockoff filter.

Constructing knockoff statistics for groups at each layer

1. First, construct group knockoff[4] variables $\widetilde{\boldsymbol{X}}^m$ satisfying

$$(\boldsymbol{X}, \widetilde{\boldsymbol{X}}^m)_{\text{swap}(\mathcal{C})} \stackrel{d}{=} (\boldsymbol{X}, \widetilde{\boldsymbol{X}}^m).$$

where $\mathcal{C}$ is any union of groups at the mth layer.

[4] Barber and Dai, 2016

Constructing knockoff statistics for groups at each layer

1. First, construct group knockoff[4] variables $\widetilde{\boldsymbol{X}}^m$ satisfying

$$(\boldsymbol{X}, \widetilde{\boldsymbol{X}}^m)_{\text{swap}(\mathcal{C})} \stackrel{d}{=} (\boldsymbol{X}, \widetilde{\boldsymbol{X}}^m).$$

   where $\mathcal{C}$ is any union of groups at the mth layer.

2. Define $(\boldsymbol{b}^\star(\lambda), \widetilde{\boldsymbol{b}}^\star(\lambda))$ via the regularized regression

$$\arg\max_{\boldsymbol{b}, \widetilde{\boldsymbol{b}}} \frac{1}{2} \left\| \boldsymbol{y} - [\boldsymbol{X}\ \widetilde{\boldsymbol{X}}^m] \begin{pmatrix} \boldsymbol{b} \\ \widetilde{\boldsymbol{b}} \end{pmatrix} \right\|^2 + \lambda \left(\sum_{g=1}^{G_m} \ell_g^m(\boldsymbol{b}_{\mathcal{A}_g^m}) + \sum_{g=1}^{G_m} \ell_g^m(\widetilde{\boldsymbol{b}}_{\mathcal{A}_g^m}) \right),$$

[4] Barber and Dai, 2016

Constructing knockoff statistics for groups at each layer

1. First, construct group knockoff[4] variables $\widetilde{\boldsymbol{X}}^m$ satisfying

$$(\boldsymbol{X}, \widetilde{\boldsymbol{X}}^m)_{\text{swap}(\mathcal{C})} \stackrel{d}{=} (\boldsymbol{X}, \widetilde{\boldsymbol{X}}^m).$$

   where $\mathcal{C}$ is any union of groups at the mth layer.

2. Define $(\boldsymbol{b}^\star(\lambda), \widetilde{\boldsymbol{b}}^\star(\lambda))$ via the regularized regression

$$\arg\max_{\boldsymbol{b}, \widetilde{\boldsymbol{b}}} \frac{1}{2} \left\| \boldsymbol{y} - [\boldsymbol{X}\ \widetilde{\boldsymbol{X}}^m] \begin{pmatrix} \boldsymbol{b} \\ \widetilde{\boldsymbol{b}} \end{pmatrix} \right\|^2 + \lambda \left(\sum_{g=1}^{G_m} \ell_g^m(\boldsymbol{b}_{\mathcal{A}_g^m}) + \sum_{g=1}^{G_m} \ell_g^m(\widetilde{\boldsymbol{b}}_{\mathcal{A}_g^m}) \right),$$

3. Let Z_g^m ($\widetilde{Z}_g^m$) be first entry times of each (knockoff) group onto the regularization path.

[4] Barber and Dai, 2016

Constructing knockoff statistics for groups at each layer

1. First, construct group knockoff[4] variables $\widetilde{\boldsymbol{X}}^m$ satisfying

$$(\boldsymbol{X}, \widetilde{\boldsymbol{X}}^m)_{\text{swap}(\mathcal{C})} \stackrel{d}{=} (\boldsymbol{X}, \widetilde{\boldsymbol{X}}^m).$$

   where $\mathcal{C}$ is any union of groups at the mth layer.

2. Define $(\boldsymbol{b}^\star(\lambda), \widetilde{\boldsymbol{b}}^\star(\lambda))$ via the regularized regression

$$\arg\max_{\boldsymbol{b}, \widetilde{\boldsymbol{b}}} \frac{1}{2} \left\| \boldsymbol{y} - [\boldsymbol{X}\ \widetilde{\boldsymbol{X}}^m] \begin{pmatrix} \boldsymbol{b} \\ \widetilde{\boldsymbol{b}} \end{pmatrix} \right\|^2 + \lambda \left(\sum_{g=1}^{G_m} \ell_g^m(\boldsymbol{b}_{\mathcal{A}_g^m}) + \sum_{g=1}^{G_m} \ell_g^m(\widetilde{\boldsymbol{b}}_{\mathcal{A}_g^m}) \right),$$

3. Let Z_g^m ($\widetilde{Z}_g^m$) be first entry times of each (knockoff) group onto the regularization path.

4. Let $W_g^m = \max(Z_g^m, \widetilde{Z}_g^m) \cdot \text{sign}(Z_g^m - \widetilde{Z}_g^m)$.

[4] Barber and Dai, 2016

Multilayer Knockoff Filter

Data: $\boldsymbol{X}$, $\boldsymbol{y}$, groupings $\{\mathcal{A}_g^m\}_{g,m}$, FDR target levels $q_1, \ldots, q_M$

Multilayer Knockoff Filter

Data: $\boldsymbol{X}$, $\boldsymbol{y}$, groupings $\{\mathcal{A}_g^m\}_{g,m}$, FDR target levels $q_1, \ldots, q_M$

for $m = 1$ **to** M **do**

  Construct group knockoff variables $\widetilde{\boldsymbol{X}}^m$;

  Construct group knockoff statistics $\boldsymbol{W}^m = (W_1^m, \ldots, W_{G_m}^m) = w^m([\boldsymbol{X}\ \widetilde{\boldsymbol{X}}^m], \boldsymbol{y})$;

end

Multilayer Knockoff Filter

Data: $\boldsymbol{X}$, $\boldsymbol{y}$, groupings $\{\mathcal{A}_g^m\}_{g,m}$, FDR target levels $q_1, \ldots, q_M$

1 **for** $m = 1$ **to** M **do**

2 | Construct group knockoff variables $\widetilde{\boldsymbol{X}}^m$;

3 | Construct group knockoff statistics $\boldsymbol{W}^m = (W_1^m, \ldots, W_{G_m}^m) = w^m([\boldsymbol{X}\ \widetilde{\boldsymbol{X}}^m], \boldsymbol{y})$;

4 **end**

5 For $\boldsymbol{t} = (t_1, \ldots, t_M)$, define $\mathcal{S}(\boldsymbol{t}) = \{j : W_{g(j,m)}^m \geq t_m\ \forall m\}$;

Multilayer Knockoff Filter

Data: $\boldsymbol{X}$, $\boldsymbol{y}$, groupings $\{\mathcal{A}_g^m\}_{g,m}$, FDR target levels $q_1, \ldots, q_M$

for $m = 1$ **to** M **do**

  Construct group knockoff variables $\widetilde{\boldsymbol{X}}^m$;

  Construct group knockoff statistics

  $\boldsymbol{W}^m = (W_1^m, \ldots, W_{G_m}^m) = w^m([\boldsymbol{X}\ \widetilde{\boldsymbol{X}}^m], \boldsymbol{y})$;

end

For $\boldsymbol{t} = (t_1, \ldots, t_M)$, define $\mathcal{S}(\boldsymbol{t}) = \{j : W^m_{g(j,m)} \geq t_m\ \forall m\}$;

For each m, define $\widehat{\mathrm{FDP}}_m(\boldsymbol{t}) = \dfrac{1 + |\{g : W_g^m \leq -t_m\}|}{|\mathcal{S}_m(\boldsymbol{t})|}$;

Multilayer Knockoff Filter

Data: $\boldsymbol{X}$, $\boldsymbol{y}$, groupings $\{\mathcal{A}_g^m\}_{g,m}$, FDR target levels $q_1, \ldots, q_M$

for $m = 1$ **to** M **do**

  Construct group knockoff variables $\widetilde{\boldsymbol{X}}^m$;

  Construct group knockoff statistics $\boldsymbol{W}^m = (W_1^m, \ldots, W_{G_m}^m) = w^m([\boldsymbol{X}\ \widetilde{\boldsymbol{X}}^m], \boldsymbol{y})$;

end

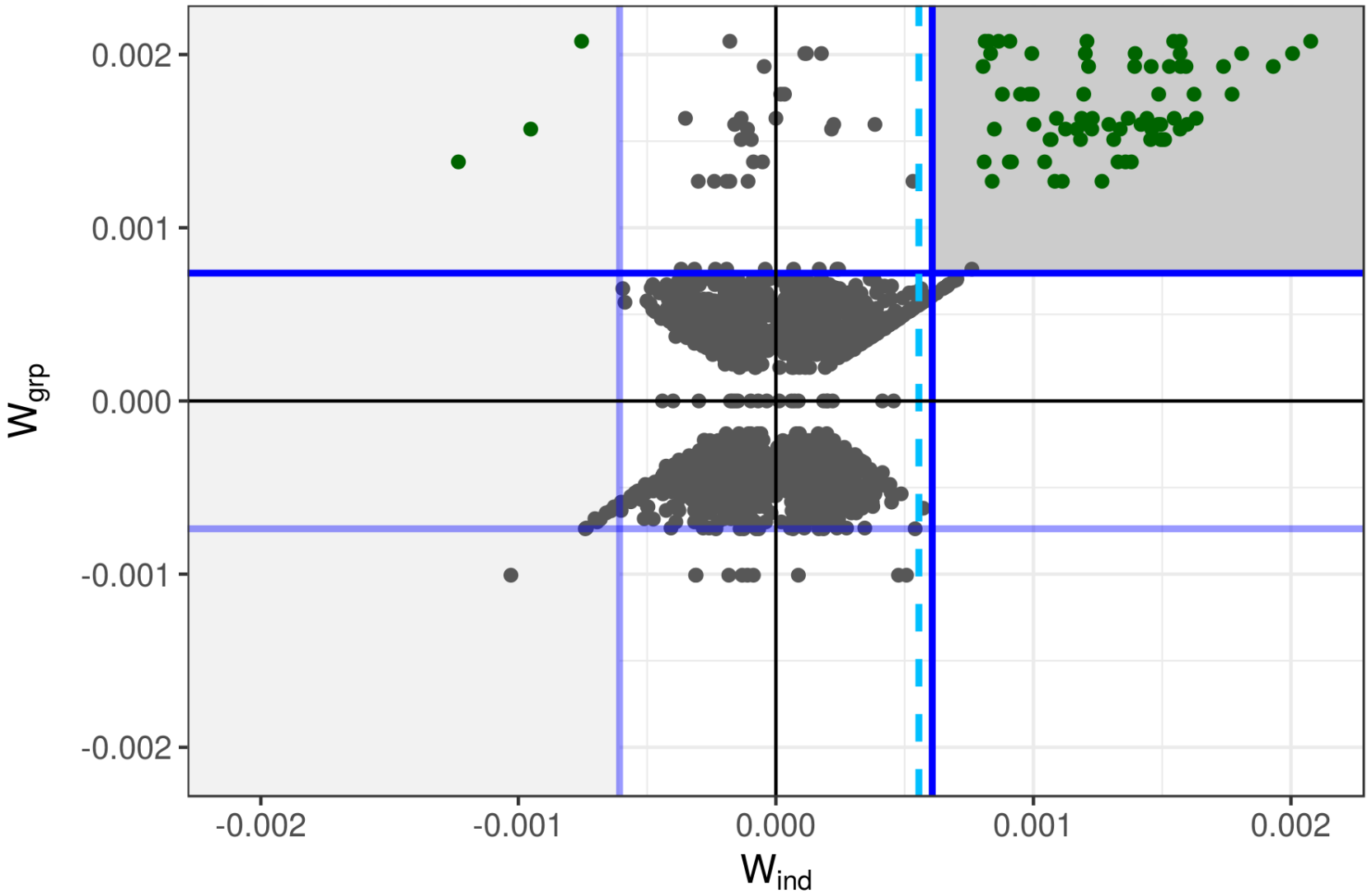
For $\boldsymbol{t} = (t_1, \ldots, t_M)$, define $\mathcal{S}(\boldsymbol{t}) = \{j : W^m_{g(j,m)} \geq t_m\ \forall m\}$;

For each m, define $\widehat{\mathrm{FDP}}_m(\boldsymbol{t}) = \dfrac{1 + |\{g : W_g^m \leq -t_m\}|}{|\mathcal{S}_m(\boldsymbol{t})|}$;

Find $\boldsymbol{t}^* = \min\{\boldsymbol{t} : \widehat{\mathrm{FDP}}_m(\boldsymbol{t}) \leq q_m\ \forall m\}$;

Result: Selection set $\mathcal{S} = \mathcal{S}(\boldsymbol{t}^*)$.

Multilayer Knockoff Filter



Multilayer FDR control

Theorem

For any valid construction of group knockoff statistics, MKF satisfies

$$FDR_m \leq c \cdot q_m,$$

where $c = 1.93$.

Generality of MKF procedure and theoretical result

Statistics $\boldsymbol{W}^m$ can have arbitrary dependencies across layers.

Pay constant factor c in theory but not in practice.

Section 4

Results on simulated and real data

Numerical simulation setup

- $n = 4500, p = 2000$
- $\boldsymbol{X}$ generated row-wise from AR(1) process with correlation ρ
- $\boldsymbol{y}$ generated from low-dimensional linear model:

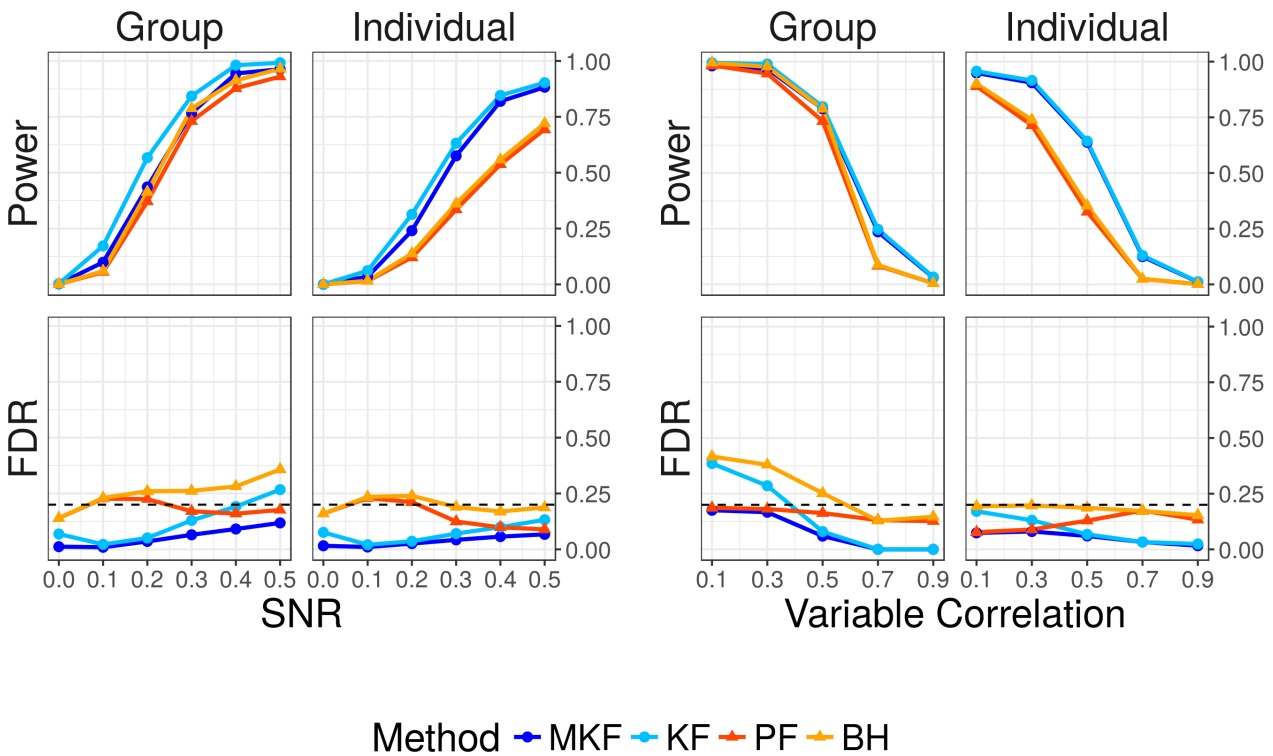
$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- Ground truth $\boldsymbol{\beta}$ has 75 non-null elements
- $M = 2$, with singleton layer and group layer
- 200 groups of size 10 each

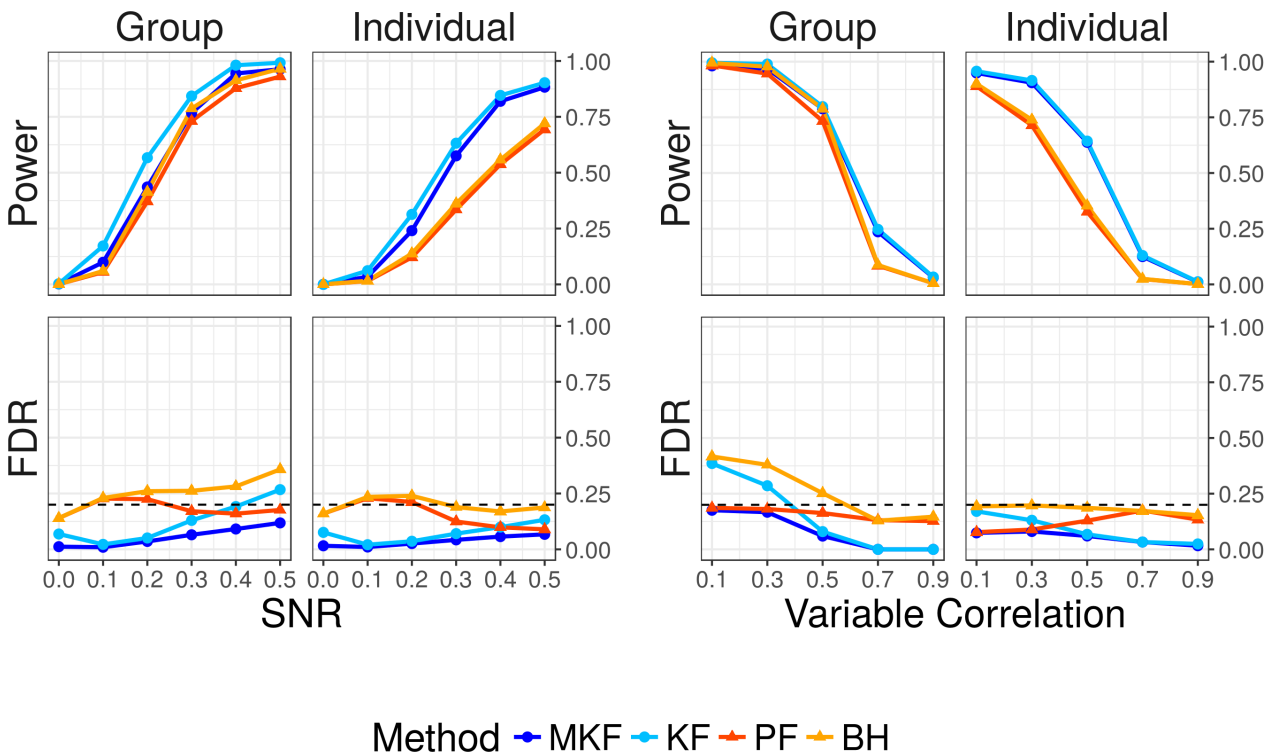
Methods compared

Method	**Multilayer?**	**Type**
Multilayer knockoff filter (MKF)	Yes	Knockoffs
Knockoff filter (KF)	No	Knockoffs
p-filter (PF)	Yes	p-values
Benjamini-Hochberg (BH)	No	p-values

Results

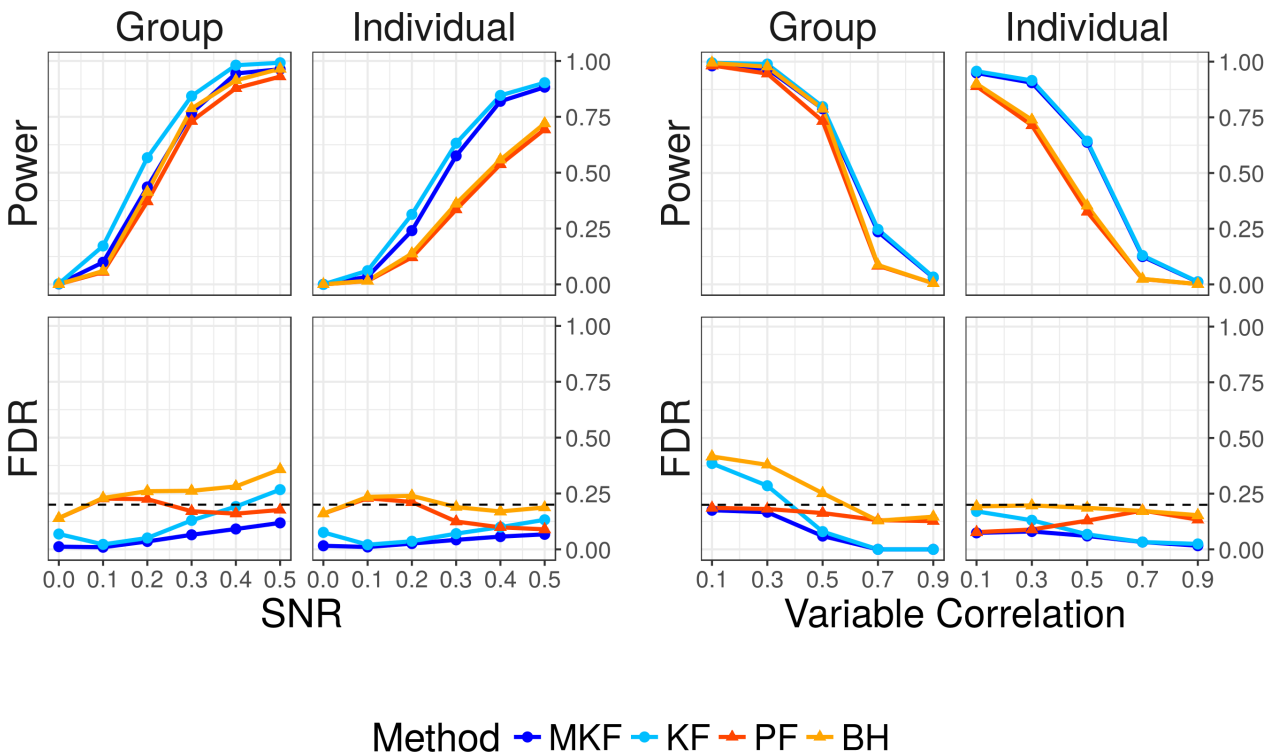

Group
Individual
Group
Individual
Power
FDR
Power
FDR
1.00
0.75
0.50
0.25
0.00
0.0 0.1 0.2 0.3 0.4 0.5
SNR
0.1 0.3 0.5 0.7 0.9
Variable Correlation
Method MKF KF PF BH

Results



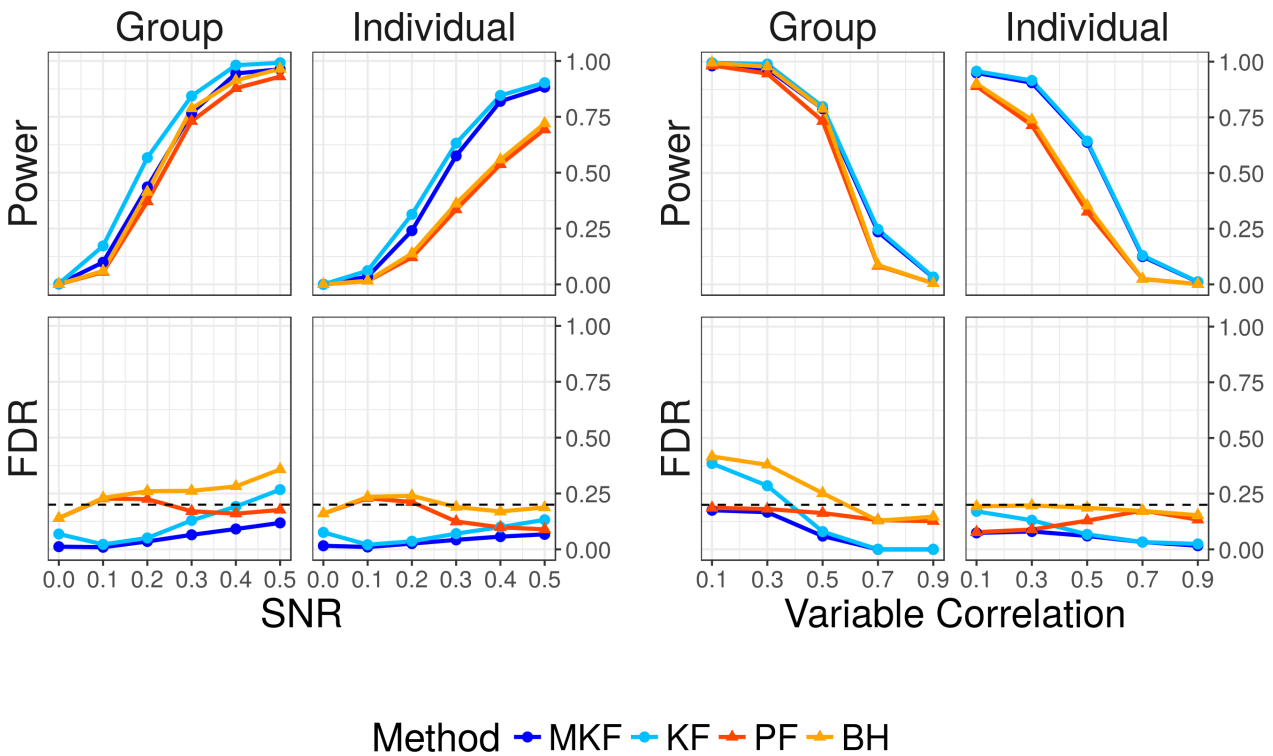
- MKF controls both FDRs

Results



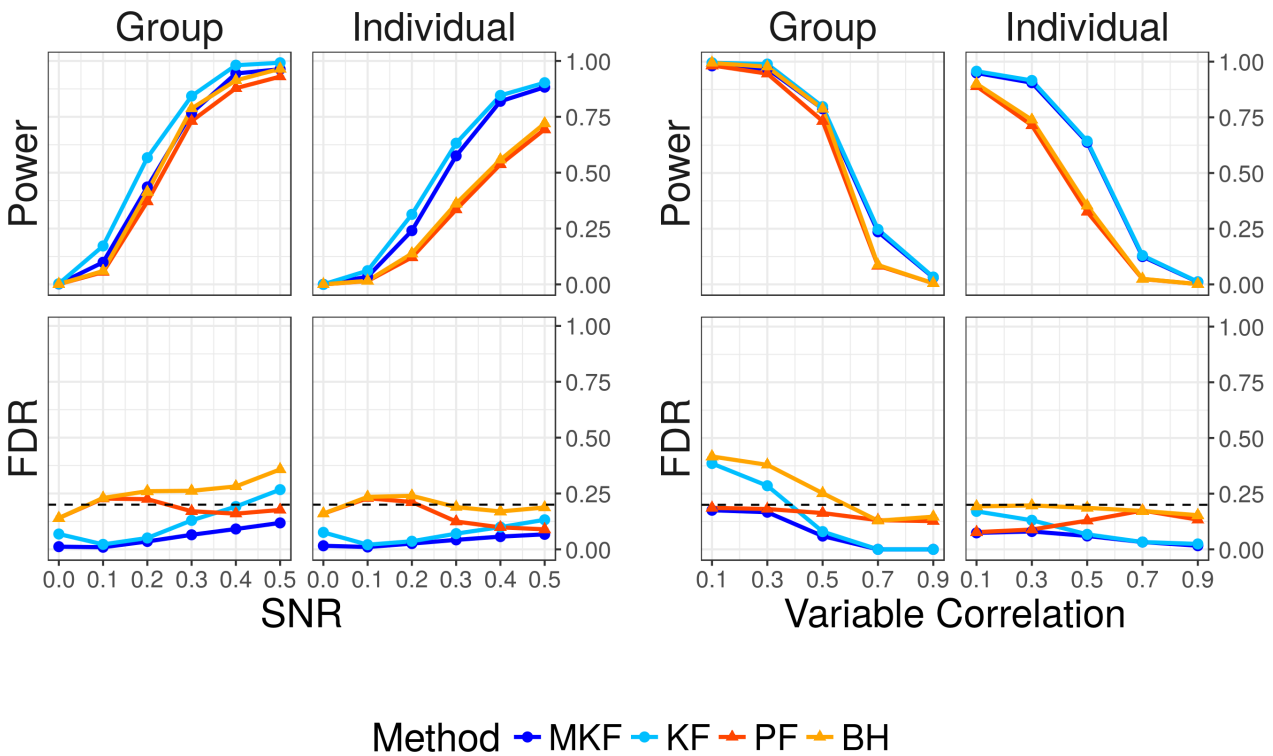
- MKF controls both FDRs
- Single-layer methods lose group FDR control

Results



- MKF controls both FDRs
- Single-layer methods lose group FDR control
- Knockoff methods more powerful than p-value methods

Results



- MKF controls both FDRs
- Single-layer methods lose group FDR control
- Knockoff methods more powerful than p-value methods
- MKF has comparable power to KF

Resequencing data for HDL cholesterol[5]

Data.

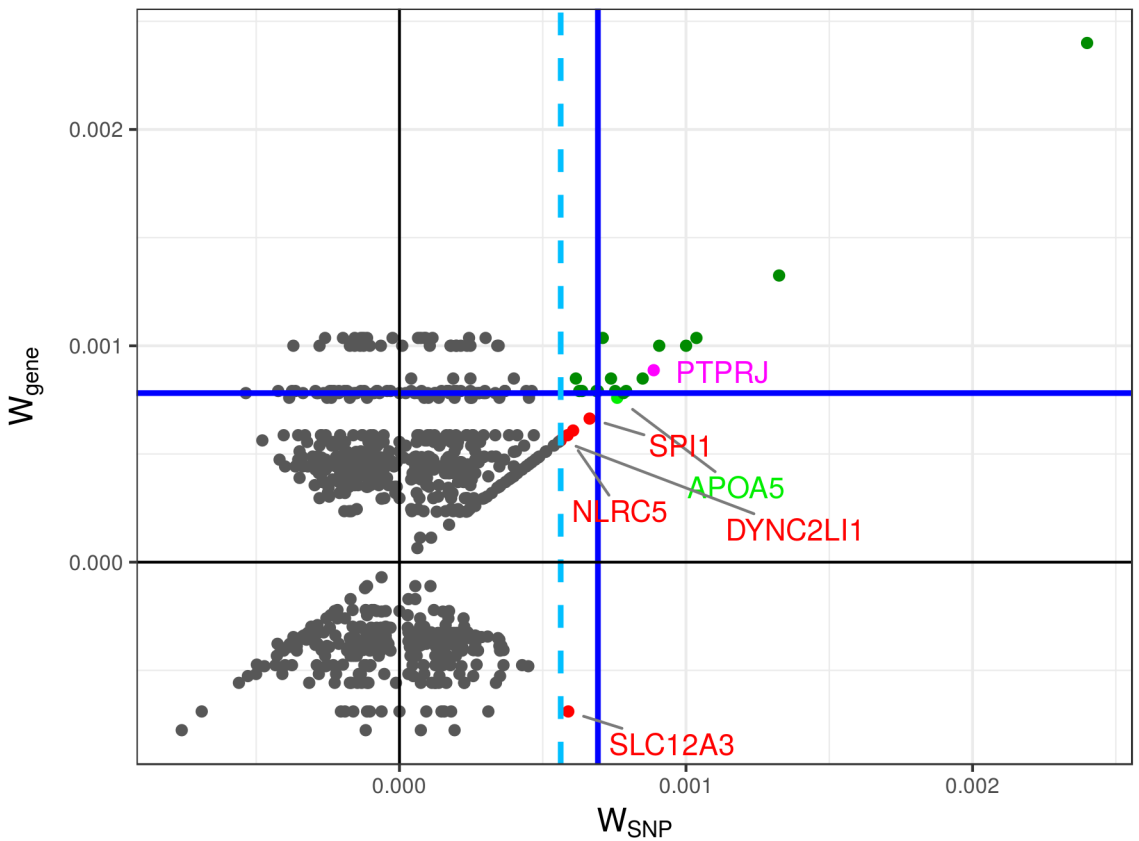
- $n = 5335$ individuals
- $p = 768$ genetic variants
- $G = 85$ genes

Methods compared.

- MKF with $q_{\text{SNP}} = q_{\text{gene}} = 0.1$.
- KF with $q_{\text{SNP}} = 0.1$.

[5] Originally analyzed in Service et. al. '14

Results on a genetic dataset



Gene	Method
ABCA1	KF, MKF
CETP	KF, MKF
GALNT2	KF, MKF
LIPC	KF, MKF
LPL	KF, MKF
PTPRJ	KF, MKF
APOA5	KF
NLRC5	KF
SLC12A3	KF
DYNC2LI1	KF
SPI1	KF

Removed four false positive genes at the cost of one false negative.

Conclusions

- For reproducibility, FDR guarantees should be provided at each layer of interpretation.

Conclusions

- For reproducibility, FDR guarantees should be provided at each layer of interpretation.
- The multilayer knockoff filter makes this possible without much power loss.

Conclusions

- For reproducibility, FDR guarantees should be provided at each layer of interpretation.
- The multilayer knockoff filter makes this possible without much power loss.
- Future work includes extension to multi-task regression and application to genome-scale data sets.

Acknowledgements

Chiara Sabatti

Emmanuel Candès

David Siegmund