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

Scientific question:

- Which single nucleotide polymorphisms (SNPs) are associated to the phenotype?
A typical GWAS output table

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

**Figure:** Source: Harold et al. Nature Genetics 41.10 (2009): 1088.
A typical GWAS output table

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

**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 \( P \leq 1 \times 10^{-5} \)

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

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 $A^m_g$:

\[ \{1, \ldots, p\} = \bigcup_{g=1}^{G_m} A^m_g. \]
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 $A_g^m$:
  \[ \{1, \ldots, p\} = \bigcup_{g=1}^{G_m} A_g^m. \]
- Selection set $S$ induces selections at each layer:
  \[ S_m = \{g = 1, \ldots, G_m : A_g^m \text{ intersects } S\}. \]
Multilayer FDR control

Definition

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

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

---

\(^1\)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\textsuperscript{2}

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

1. Get group p-values using Simes.

\textsuperscript{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 $t = (t_1, \ldots, t_M)$.

---

$^2$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 $t = (t_1, \ldots, t_M)$.
3. Select hypotheses $S(t)$ passing thresholds at all layers.

---

$^2$Barber and Ramdas ‘15
p-filter\textsuperscript{2}

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

1. Get group p-values using Simes.
2. Introduce thresholds $t = (t_1, \ldots, t_M)$.
3. Select hypotheses $S(t)$ passing thresholds at all layers.
4. Choose $t^* = \max \{ t : \widehat{\text{FDP}}_m(t) \leq q_m \ \forall m \}$. 

\textsuperscript{2}Barber and Ramdas ‘15
Knockoff filter\textsuperscript{3}

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

\textsuperscript{3}Barber and Candes '15
Knockoff filter\textsuperscript{3}

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

1. Construct “knockoff variables” $\tilde{X}$ to use as controls.

\textsuperscript{3}Barber and Candes ‘15
Knockoff filter$^3$

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

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

$^3$Barber and Candes ‘15
Knockoff filter³

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

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

³Barber and Candes ‘15
Knockoff filter\(^3\)

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

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

\(^3\)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 \(\tilde{X}^m\) satisfying

\[
(X, \tilde{X}^m)_{\text{swap}(C)} \overset{d}{=} (X, \tilde{X}^m).
\]

where \(C\) is any union of groups at the \(m\)th layer.

\(^4\)Barber and Dai, 2016
Constructing knockoff statistics for groups at each layer

1. First, construct group knockoff\footnote{Barber and Dai, 2016} variables $\tilde{X}^m$ satisfying

$$(X, \tilde{X}^m)_{\text{swap}(C)} \overset{d}{=} (X, \tilde{X}^m).$$

where $C$ is any union of groups at the $m$th layer.

2. Define $(b^*(\lambda), \tilde{b}^*(\lambda))$ via the regularized regression

$$\arg\max_{b, \tilde{b}} \frac{1}{2} \left\| y - [X \ \tilde{X}^m] \begin{pmatrix} b \\ \tilde{b} \end{pmatrix} \right\|^2 + \lambda \left( \sum_{g=1}^{G_m} \ell^m_g(b_{A_g}) + \sum_{g=1}^{G_m} \ell^m_g(\tilde{b}_{A_g}) \right),$$

where $\ell^m_g$ is a loss function for group $g$. 
Constructing knockoff statistics for groups at each layer

1. First, construct group knockoff\(^4\) variables \(\tilde{X}^m\) satisfying

\[
(X, \tilde{X}^m)_{\text{swap}(C)} \overset{d}{=} (X, \tilde{X}^m).
\]

where \(C\) is any union of groups at the \(m\)th layer.

2. Define \((b^*(\lambda), \tilde{b}^*(\lambda))\) via the regularized regression

\[
\arg\max_{b, \tilde{b}} \frac{1}{2} \left\| y - [X \tilde{X}^m] \begin{pmatrix} b \\ \tilde{b} \end{pmatrix} \right\|^2 + \lambda \left( \sum_{g=1}^{G_m} \ell_g^m(b_{A_g}) + \sum_{g=1}^{G_m} \ell_g^m(\tilde{b}_{A_g}) \right),
\]

3. Let \(Z_g^m (\tilde{Z}_g^m)\) 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 variables $\tilde{X}^m$ satisfying

$$\begin{align*}
(X, \tilde{X}^m)_{\text{swap}(C)} &\overset{d}{=} (X, \tilde{X}^m).
\end{align*}$$

where $C$ is any union of groups at the $m$th layer.

2. Define $(b^*(\lambda), \tilde{b}^*(\lambda))$ via the regularized regression

$$\arg\max_{\tilde{b}, \tilde{b}} \frac{1}{2} \left\| y - [X \tilde{X}^m] \begin{pmatrix} \tilde{b} \\ \tilde{b} \end{pmatrix} \right\|^2 + \lambda \left( \sum_{g=1}^{G_m} \ell_m^g(b_{A^m_g}) + \sum_{g=1}^{G_m} \ell_m^g(\tilde{b}_{A^m_g}) \right),$$

3. Let $Z^m_g (\tilde{Z}^m_g)$ be first entry times of each (knockoff) group onto the regularization path.

4. Let $W^m_g = \max(Z^m_g, \tilde{Z}^m_g) \cdot \text{sign}(Z^m_g - \tilde{Z}^m_g)$.

---

$^4$Barber and Dai, 2016
Multilayer Knockoff Filter

**Data:** $X$, $y$, groupings $\{A^m_g\}_{g,m}$, FDR target levels $q_1, \ldots, q_M$
Multilayer Knockoff Filter

**Data:** \(X, y,\) groupings \(\{A_g^m\}_{g,m},\) FDR target levels \(q_1, \ldots, q_M\)

1. **for** \(m = 1 \text{ to } M\) **do**
2. Construct group knockoff variables \(\tilde{X}^m;\)
3. Construct group knockoff statistics
   \[W^m = (W_1^m, \ldots, W_{G_m}^m) = w^m([X \; \tilde{X}^m], y);\]
4. **end**
Multilayer Knockoff Filter

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

1. for $m = 1$ to $M$ do
2.   Construct group knockoff variables $\tilde{X}^m$;
3.   Construct group knockoff statistics
   \[ W^m = (W^m_1, \ldots, W^m_{G_m}) = w^m([X \ \tilde{X}^m], y); \]
4. end
5. For $t = (t_1, \ldots, t_M)$, define $S(t) = \{j : W^m_{g(j,m)} \geq t_m \ \forall m\}$;
Multilayer Knockoff Filter

**Data:** $X$, $y$, groupings $\{A_g^m\}_{g,m}$, FDR target levels $q_1, \ldots, q_M$

1. **for** $m = 1$ **to** $M$ **do**
   2. Construct group knockoff variables $\tilde{X}^m$;
   3. Construct group knockoff statistics
      $$W^m = (W_1^m, \ldots, W_{G_m}^m) = w^m([X \tilde{X}^m], y);$$
4. **end**
5. For $t = (t_1, \ldots, t_M)$, define
   $$S(t) = \{j : W_{g(j,m)}^m \geq t_m \ \forall m\};$$
6. For each $m$, define
   $$\widehat{\text{FDP}}_m(t) = \frac{1 + |\{g : W_{g}^m \leq -t_m\}|}{|S_m(t)|};$$
Multilayer Knockoff Filter

**Data:** \( X, y, \) groupings \( \{A_g^m\}_{g,m}, \) FDR target levels \( q_1, \ldots, q_M \)

1. **for** \( m = 1 \) **to** \( M \) **do**
   2. Construct group knockoff variables \( \tilde{X}^m \);
   3. Construct group knockoff statistics
      \[
      W^m = (W_1^m, \ldots, W_{G_m}^m) = w^m([X \ \tilde{X}^m], y);
      \]
   4. **end**

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

6. For each \( m \), define \( \widehat{\text{FDP}}_m(t) = \frac{1 + |\{g : W_g^m \leq -t_m\}|}{|S_m(t)|} \);

7. Find \( t^* = \min\{t : \widehat{\text{FDP}}_m(t) \leq q_m \ \forall m\} \);

**Result:** Selection set \( S = S(t^*) \).
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 $W^m$ 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

- $n = 4500, p = 2000$
- $\mathbf{X}$ generated row-wise from AR(1) process with correlation $\rho$
- $\mathbf{y}$ generated from low-dimensional linear model:
  \[ \mathbf{y} = \mathbf{X}\beta + \epsilon \]
- Ground truth $\beta$ has 75 non-null elements
- $M = 2$, with singleton layer and group layer
- 200 groups of size 10 each
## Methods compared

<table>
<thead>
<tr>
<th>Method</th>
<th>Multilayer?</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multilayer knockoff filter (MKF)</td>
<td>Yes</td>
<td>Knockoffs</td>
</tr>
<tr>
<td>Knockoff filter (KF)</td>
<td>No</td>
<td>Knockoffs</td>
</tr>
<tr>
<td>p-filter (PF)</td>
<td>Yes</td>
<td>p-values</td>
</tr>
<tr>
<td>Benjamini-Hochberg (BH)</td>
<td>No</td>
<td>p-values</td>
</tr>
</tbody>
</table>
Results

- MKF controls both FDRs.
- Single-layer methods lose group FDR control.
- Knockoff methods are more powerful than p-value methods.
- MKF has comparable power to KF.
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\textsuperscript{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 $q_{\text{SNP}} = 0.1$.

\textsuperscript{5}Originally analyzed in Service et. al. ‘14
## Results on a genetic dataset

<table>
<thead>
<tr>
<th>Gene</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABCA1</td>
<td>KF, MKF</td>
</tr>
<tr>
<td>CETP</td>
<td>KF, MKF</td>
</tr>
<tr>
<td>GALNT2</td>
<td>KF, MKF</td>
</tr>
<tr>
<td>LIPC</td>
<td>KF, MKF</td>
</tr>
<tr>
<td>LPL</td>
<td>KF, MKF</td>
</tr>
<tr>
<td>PTPRJ</td>
<td>KF, MKF</td>
</tr>
<tr>
<td>APOA5</td>
<td>KF</td>
</tr>
<tr>
<td>NLRC5</td>
<td>KF</td>
</tr>
<tr>
<td>SLC12A3</td>
<td>KF</td>
</tr>
<tr>
<td>DYNC2LI1</td>
<td>KF</td>
</tr>
<tr>
<td>SPI1</td>
<td>KF</td>
</tr>
</tbody>
</table>

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