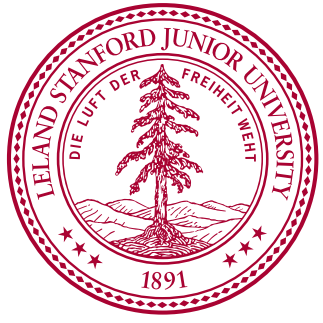
Knockoff genotypes: beauty in counterfeit

Chiara Sabatti

Biomedical Data Science & Statistics

January 10, 2018



Acknowledgements

Matteo Sesia

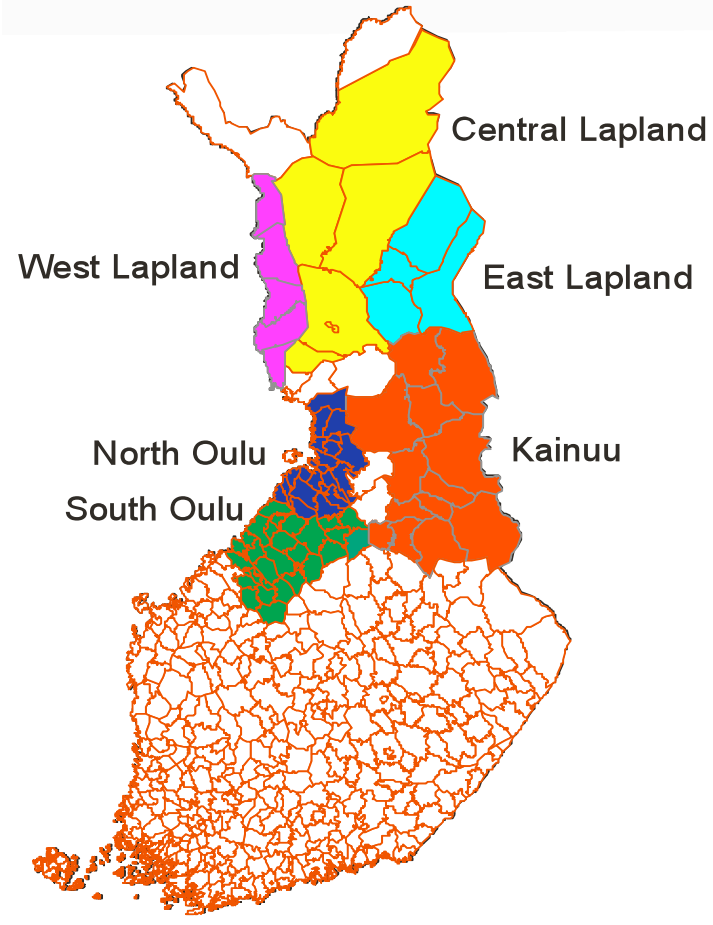
Emmanuel Candes

(and the work of many others, especially Rina Barber and Lucas Janson)

Thanks to NSF for support

An early GWAS for complex traits: lipids in NFBC66

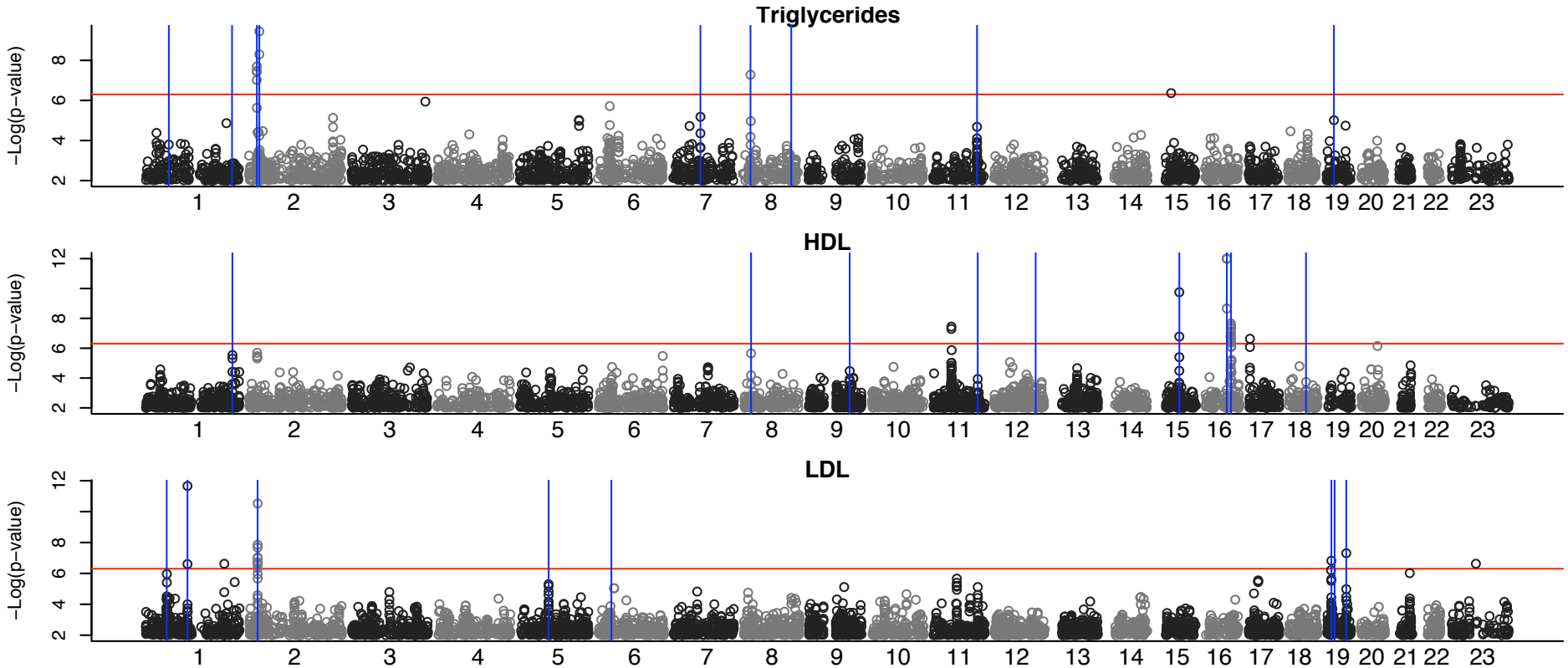
- Cohort study based in northern Finland
- Fasting serum concentrations of lipids (triglycerides, TG, high density lipoproteins, HDL, and low density lipoproteins, LDL) for $\approx$ 5400 subjects
- Genotypes at $\approx$ 300,000 Single Nucleotide Polymorphisms (SNP)



Results for blood lipids in NFBC

Test the hypothesis $H_0 : \beta_k = 0$, for each of the M SNPs, one SNP at the time,

$$y_i = \beta_0 + \beta_k X_{ik} + \text{Covariates}_i + \eta_i$$



Marginal association

Pros

- Computationally simple
- Deals easily with missing data
- p-values available
- Fairly immune to linkage disequilibrium

Cons

- Lower power due to overestimate of the error size
- "Association is not causation"
- Challenges for genetic counseling
- Challenges in interpretations across ethnic groups

Another approach: conditional testing

Null variable

Say $j \in \mathcal{H}_0$ is null iff $Y \perp\!\!\!\perp X_j \,|\, X_{-j}$

- In multivariate regression models (polygenic model), the coefficients of each variable capture this type of effect
- In the logistic model $\mathbb{P}(Y = 0|X) = \dfrac{1}{1 + e^{X^\top \beta}}$, as long as the variables $X_{1:p}$ are not perfectly dependent, then $j \in \mathcal{H}_0 \iff \beta_j = 0$
- The notion does not require specifying a form of dependence between Y and X
- Related to Markov blanket in causal literature.

Selecting the important variants with reproducibility guarantees

Question

Through which variables does the dstribution of $Y \mid X$ depends on X?

Goal

Select set $\hat{\mathcal{S}}$ of variants X_j that are likely to be relevant
without too many false positives

One way of operationalize this is to try to control false discovery rate (FDR)

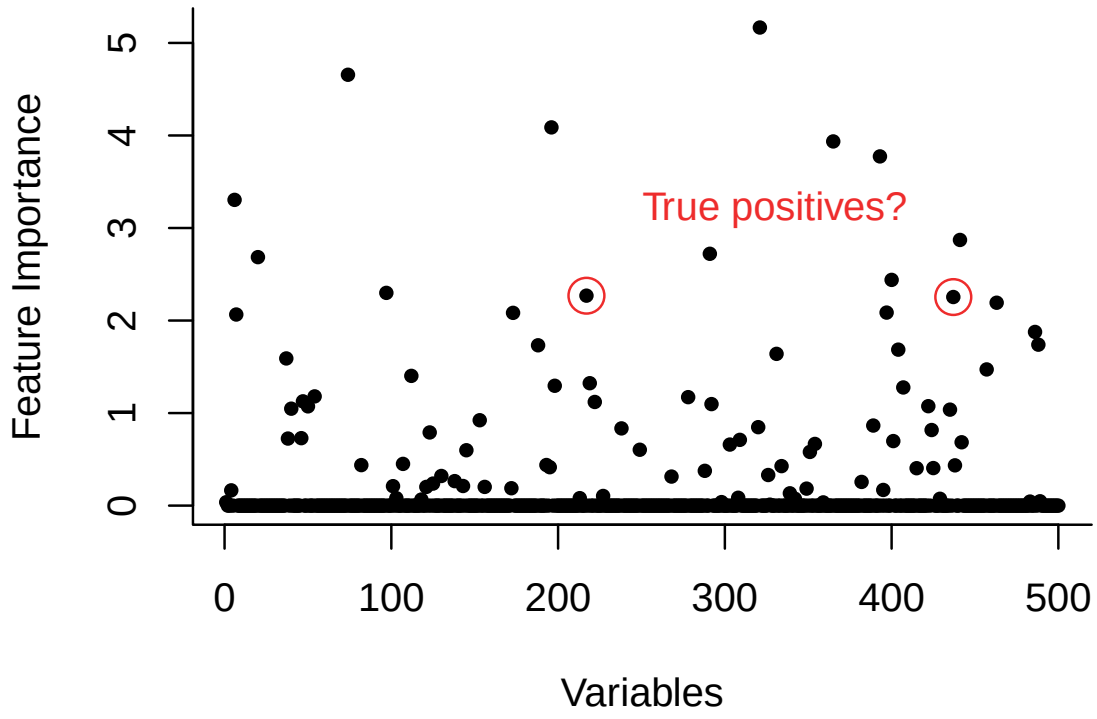
$$\text{FDR} = \mathbb{E}\,\frac{\#\text{ false positives}}{\#\text{ features selected}}$$

How can we do this?

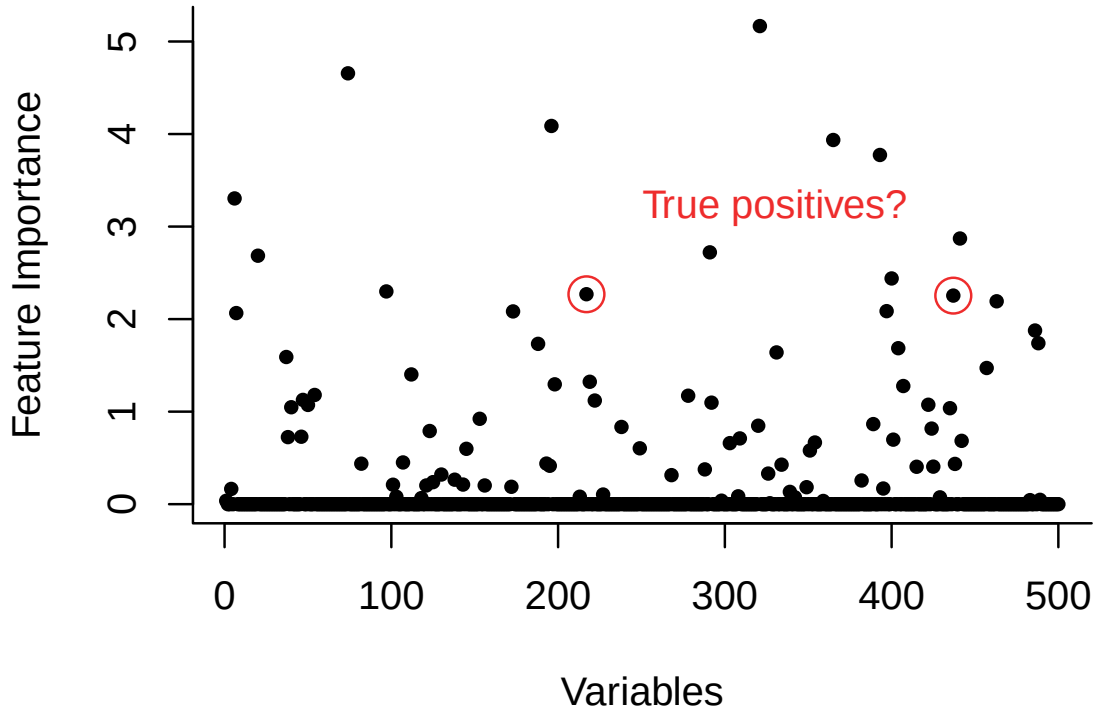
We are typically in a setting where n (number of observations) is smaller than p (number of genetic variants queried).

- If we are dealing with a quantitative trait, we might think of using the Lasso
- Lasso-like procedures exist also for binary traits
- There are a number of other approaches to selection — trees, forests, etc..

But how to decide which variables to select so that we can control FDR?



The Knockoffs framework



- It would be good to know how the feature importance statistics for the null variables looks like
- Barber and Candès (2014) introduced the idea of knockoff: variables $\tilde{X}$ that "look like" X, but are by construction independent from Y.
- Further developments Candès, Fan, Janson and Lv (2016); Katsevich and Sabatti (2017), Sesia, Sabatti Candès (2017); etc...

An idea for artificial null variables: model-X knockoffs

i.i.d. samples $(X^{(i)}, Y^{(i)}) \sim F_{XY}$

- Distribution of X known
- Distribution of $Y \,|\, X$ (likelihood) completely unknown

- Originals $X = (X_1, \ldots, X_p)$
- Knockoffs $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_p)$

(1) Pairwise exchangeability

$$(X, \tilde{X})_{\text{swap}(S)} \overset{d}{=} (X, \tilde{X})$$

e.g.

$$(X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3)_{\text{swap}(\{2,3\})} \overset{d}{=} (X_1, \tilde{X}_2, \tilde{X}_3, \tilde{X}_1, X_2, X_3)$$

(2) $\tilde{X} \perp\!\!\!\perp Y \,|\, X$ (ignore Y when constructing knockoffs)

The idea of using dummy variables is not new

For linear models, Miller ('84, '02) creates 'dummy' variables with entries drawn i.i.d. at random

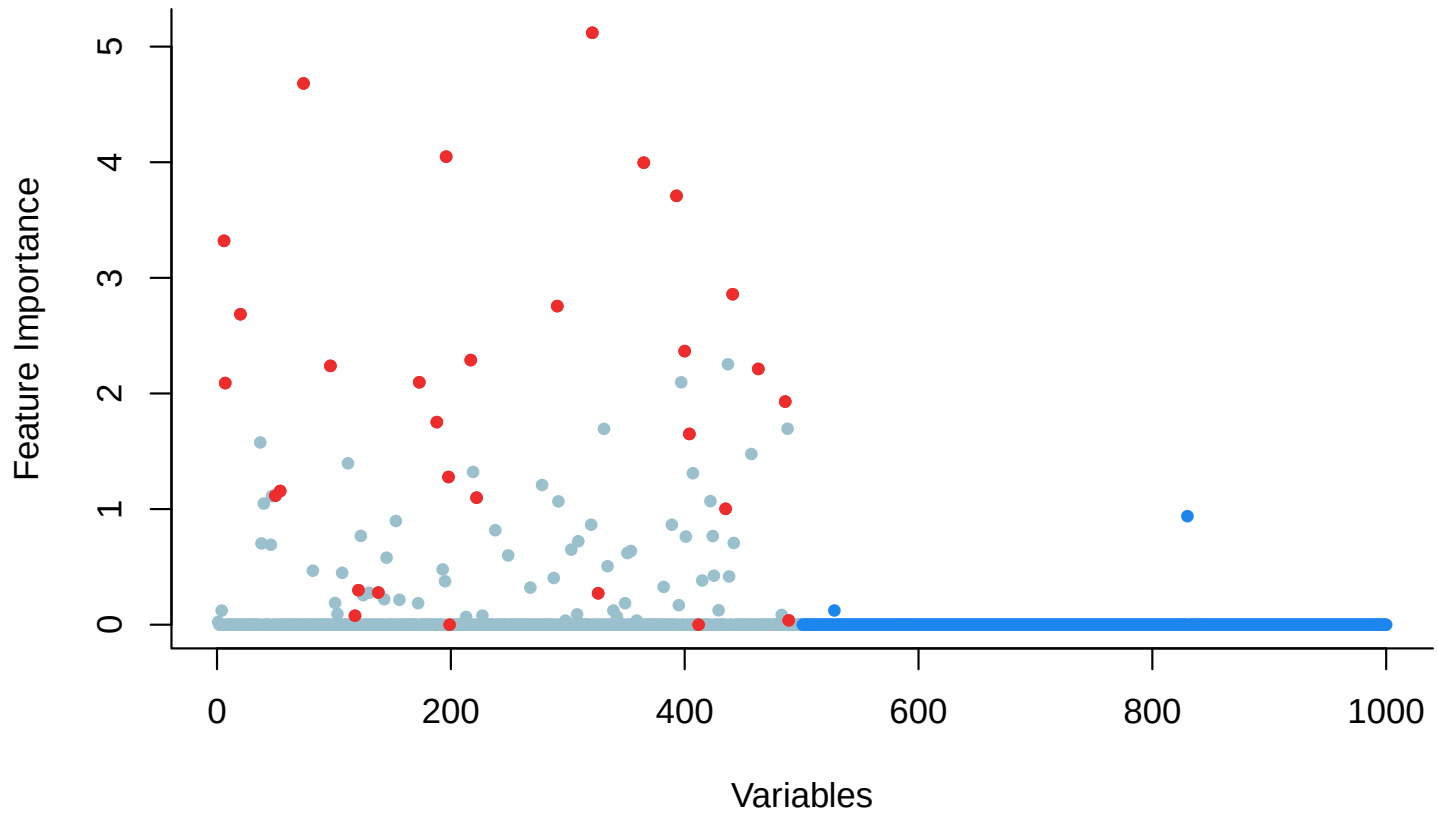
- Forward selection procedure is applied to augmented list of variables
- Stop when selects a dummy variable for the first time

Pseudovariables (permuted rows and variants): Wu, Boos and Stefanski ('07, '09)

Dummies	Structure preserved
Independent Gaussian variables	Mean and marginal variance
Permuted rows `X[sample(n),]`	(Joint) distribution
Knockoffs	More...

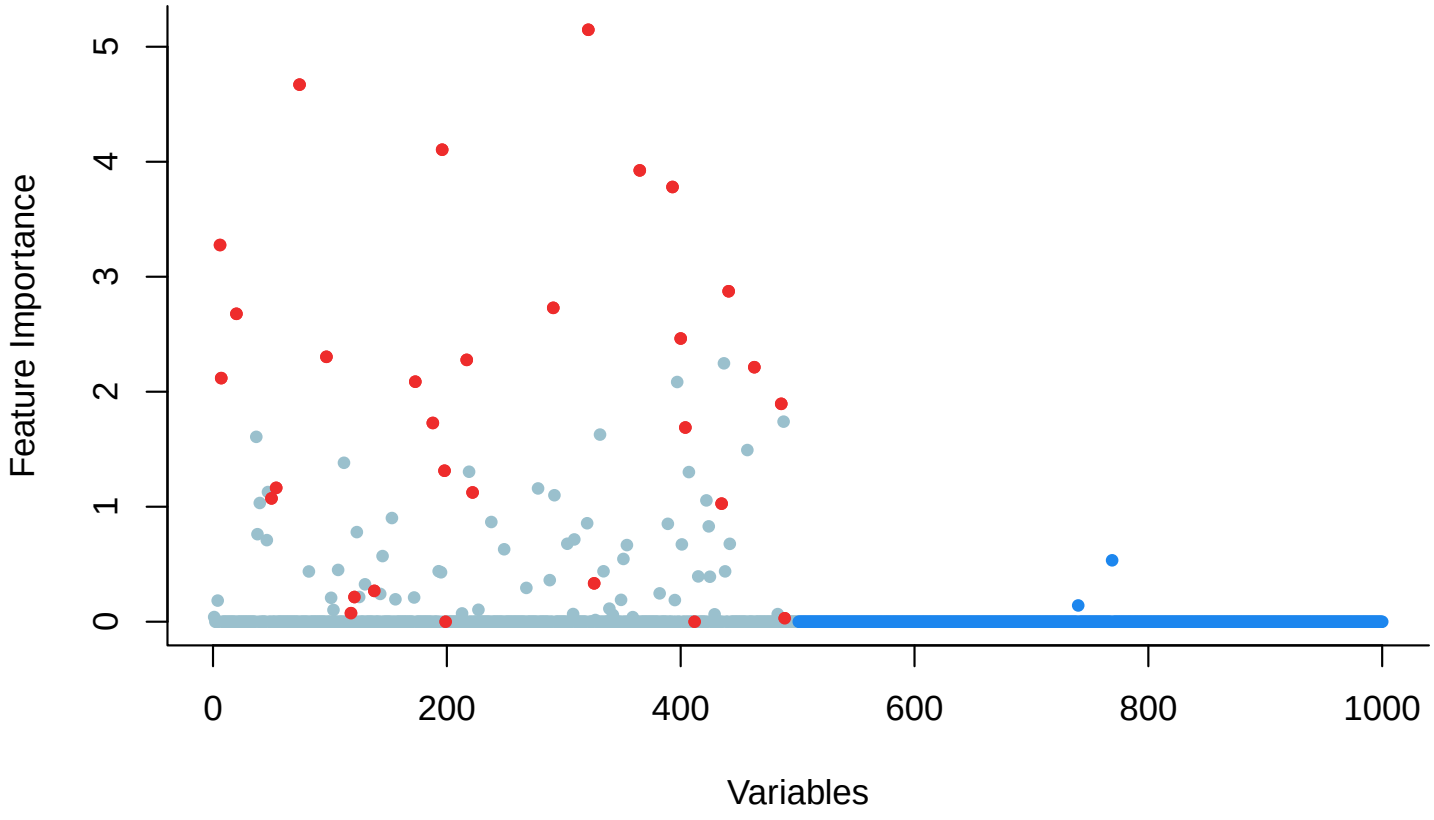
Gaussian dummies

Feature importance statistic $Z_j = |\hat{\beta}_j(\lambda = 3)|$



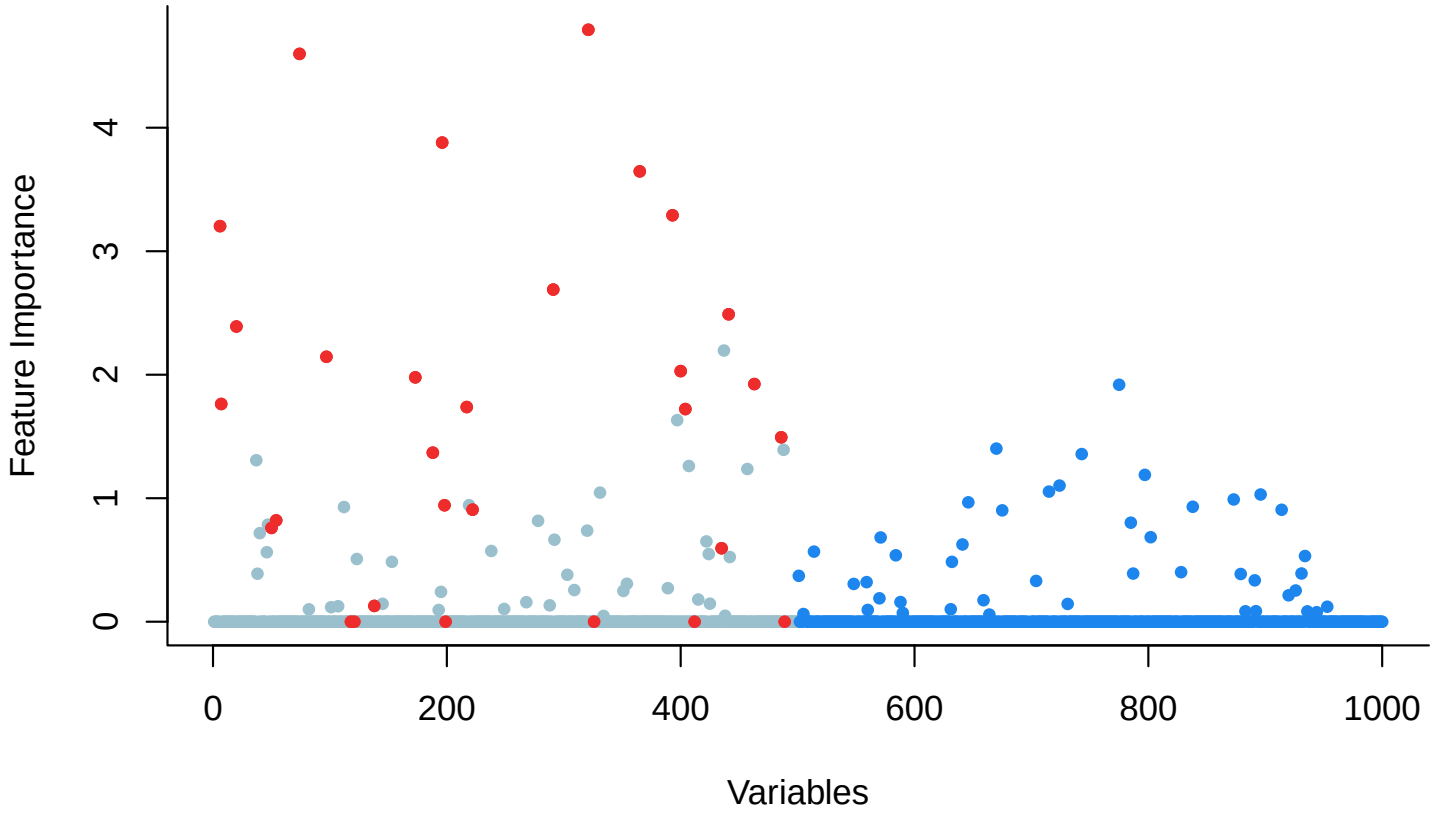
Permuted dummies

Feature importance $Z_j = |\hat{\beta}_j(\lambda = 3)|$



Knockoff dummies

Feature importance $Z_j = |\hat{\beta}_j(\lambda = 3)|$



Knockoffs do well because...

... the feature importance statistics are exchangeable

$$(\underbrace{Z_1, \ldots, Z_p}_{\text{originals}}, \underbrace{\tilde{Z}_1, \ldots, \tilde{Z}_p}_{\text{knockoffs}}) = z([\boldsymbol{X}, \tilde{\boldsymbol{X}}], \boldsymbol{y})$$

Swapping originals and knockoffs swaps the Z's

$$\overbrace{(Z_1, \tilde{Z}_2, \tilde{Z}_3, \tilde{Z}_1, Z_2, Z_3)}^{(Z, \tilde{Z})_{\text{swap}\{2,3\}}} = z([\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{\text{swap}\{2,3\}}, \boldsymbol{y})$$

Theorem (Candes, Fan, Janson and Lv)

No matter the relationship between Y and X:

$$j \in \mathcal{H}_0 \implies (Z_j, \tilde{Z}_j) \stackrel{d}{=} (\tilde{Z}_j, Z_j)$$

$$\text{more generally} \quad \mathcal{T} \subset \mathcal{H}_0 \implies (Z, \tilde{Z})_{\text{swap}(\mathcal{T})} \stackrel{d}{=} (Z, \tilde{Z})$$

How can we use the knockoffs? a little detour

1. Construct knockoff adjusted scores

Adjusted scores W_j with flip-sign property

Combine Z_j and $\tilde{Z}_j$ into single (knockoff) score W_j

$$W_j = w_j(Z_j, \tilde{Z}_j) \qquad w_j(\tilde{Z}_j, Z_j) = -w_j(Z_j, \tilde{Z}_j)$$

e.g. $\qquad W_j = Z_j - \tilde{Z}_j \qquad W_j = Z_j \vee \tilde{Z}_j \cdot \begin{cases} +1 & Z_j > \tilde{Z}_j \\ -1 & \text{else} \end{cases}$

- Large W_j says that variable j appears important
- A negative W_j says that the knockoff of variable j seams more important than the original variable
- Null W_j's are symmetrically distributed
- Conditional on $|W|$, signs of null W_j's are i.i.d. coin flips

2. Estimate the FDR

Interested in selecting $\{j : W_j \geq t\}$

$$\mathsf{FDP}(t) = \frac{\#\{j \text{ null } : W_j \geq t\}}{\#\{j : W_j \geq t\} \vee 1} \approx \frac{\#\{j \text{ null } : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1}$$

$$\leq \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} := \widehat{\mathsf{FDP}}(t)$$

3. Select the max number of variables while $\widehat{\text{FDP}} \leq q$

$$\mathcal{S}^{\pm}(t) = \{j : |W_j| \geq t \text{ and } \operatorname{sgn}(W_j) = \pm\}$$

$$\tau = \min\left\{t : \widehat{\text{FDP}}(t) = \frac{1+|\mathcal{S}^{-}(t)|}{1 \vee |\mathcal{S}^{+}(t)|} \leq q\right\}$$

$$\hat{\mathcal{S}} = \{W_j \geq \tau\}$$

Theorem (Barber and Candès ('15))

$$\mathbb{E}\left[\frac{\#\text{ false positives}}{\#\text{ selections} + q^{-1}}\right] \leq q$$

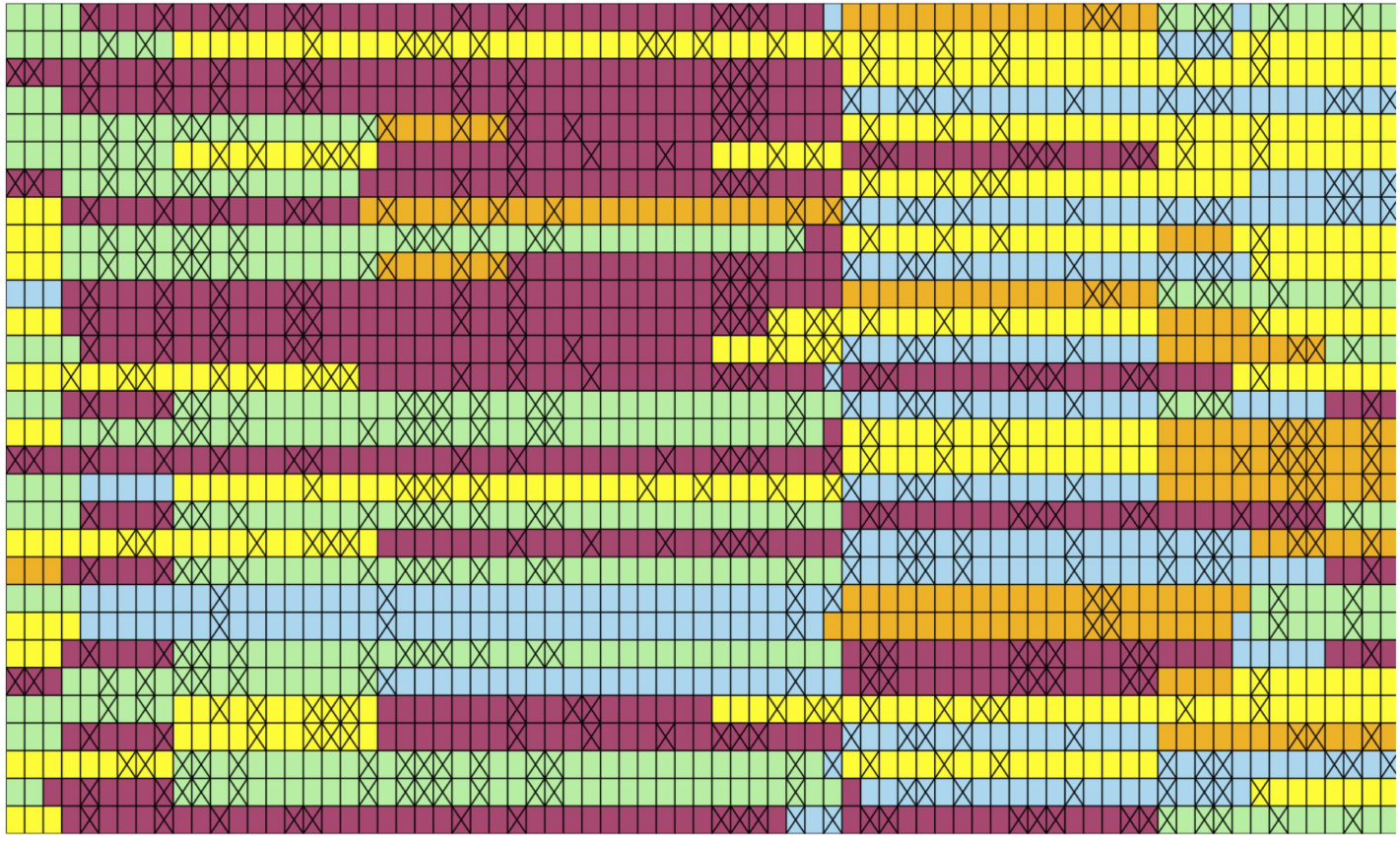
Going back to GWAS

How does the Model-X knockoff framework fit GWAS?

- i.i.d. samples $(X^{(i)}, Y^{(i)}) \sim F_{XY}$
  $\longrightarrow$ This is a good description of population samples
- Distribution of X known $\longrightarrow$ We do have a large collection of genotype data, irrespectively of phenotypes, that can be leveraged.
- Distribution of $Y \mid X$ (likelihood) completely unknown $\longrightarrow$ it is nice not to have to make assumptions here

To deploy it we need a distribution for genotypes and a method to generate knockoffs with the right exchangeability property.

A phenomenological HMM for haplotype & genotype data



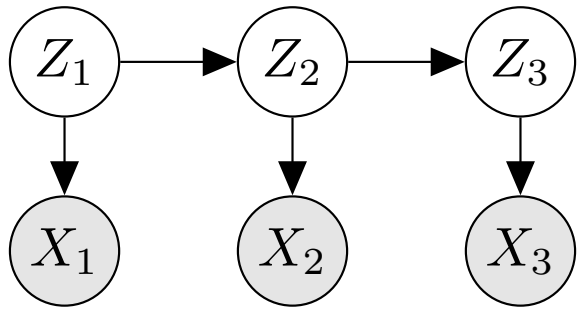
- `fastPHASE` (Scheet, '06)

- `IMPUTE` (Marchini, '07)
- `MaCH` (Li, '10)

Haplotypes as Hidden Markov Models (HMMs)

$\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ is a HMM if

$$\begin{cases} \boldsymbol{Z} \sim \text{MC}\,(q_1, \boldsymbol{Q}) & \text{(latent Markov chain)} \\ X_j|\boldsymbol{Z} \sim X_j|Z_j \overset{\text{ind.}}{\sim} f_j(X_j; Z_j) & \text{(emission distribution)} \end{cases}$$



The $\boldsymbol{Z}$ variables are latent and only the $\boldsymbol{X}$ variables are observed

A general recipe for knockoffs

Algorithm Sequential Conditional Independent Pairs

for $j = \{1, \ldots, p\}$ **do**
  Sample $\tilde{X}_j$ from law of $X_j \,|\, X_{-j}, \tilde{X}_{1:j-1}$
end

e.g. $p = 3$

- Sample $\tilde{X}_1$ from $X_1 \,|\, X_{-1}$
- Joint law of $X, \tilde{X}_1$ is known
- Sample $\tilde{X}_2$ from $X_2 \,|\, X_{-2}, \tilde{X}_1$
- Joint law of $X, \tilde{X}_{1:2}$ is known
- Sample $\tilde{X}_3$ from $X_3 \,|\, X_{-3}, \tilde{X}_{1:2}$

Joint law of $X, \tilde{X}$ is pairwise exchangeable!

Sesia, Sabatti, Candès (2017)

Usually not practical, but extremely efficient for Markov chains

Knockoff copies of a hidden Markov model

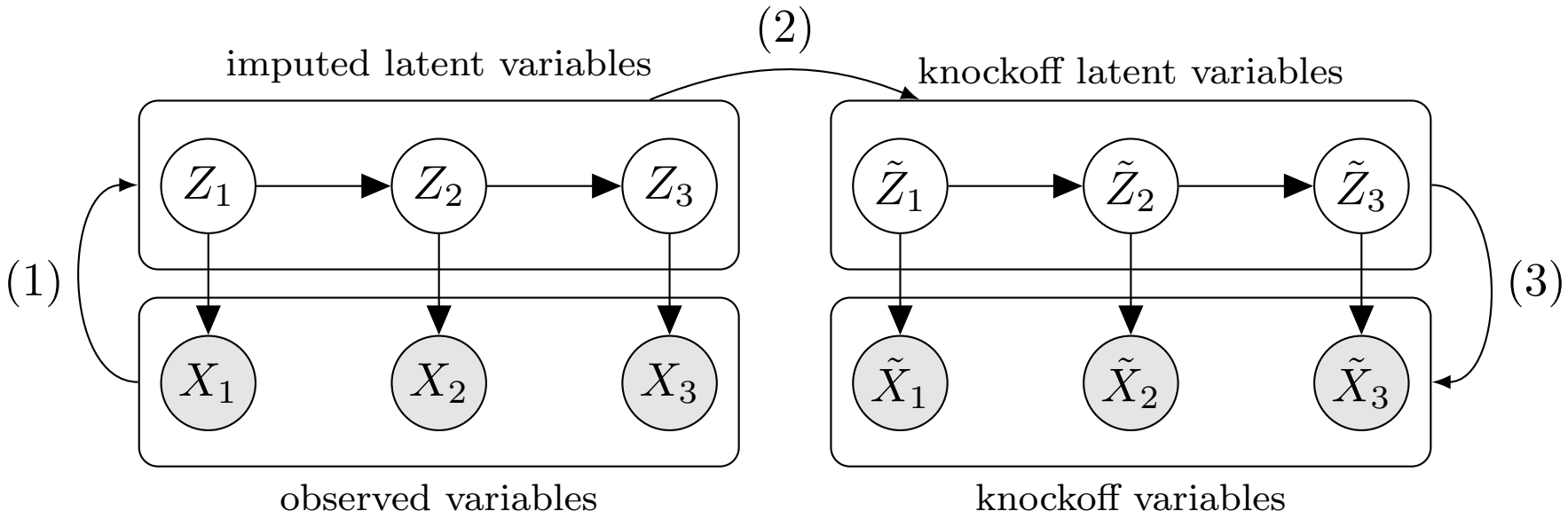
Theorem (Sesia, Sabatti, Candès '17)

A knockoff copy of $\tilde{\boldsymbol{X}}$ of $\boldsymbol{X}$ can be constructed as

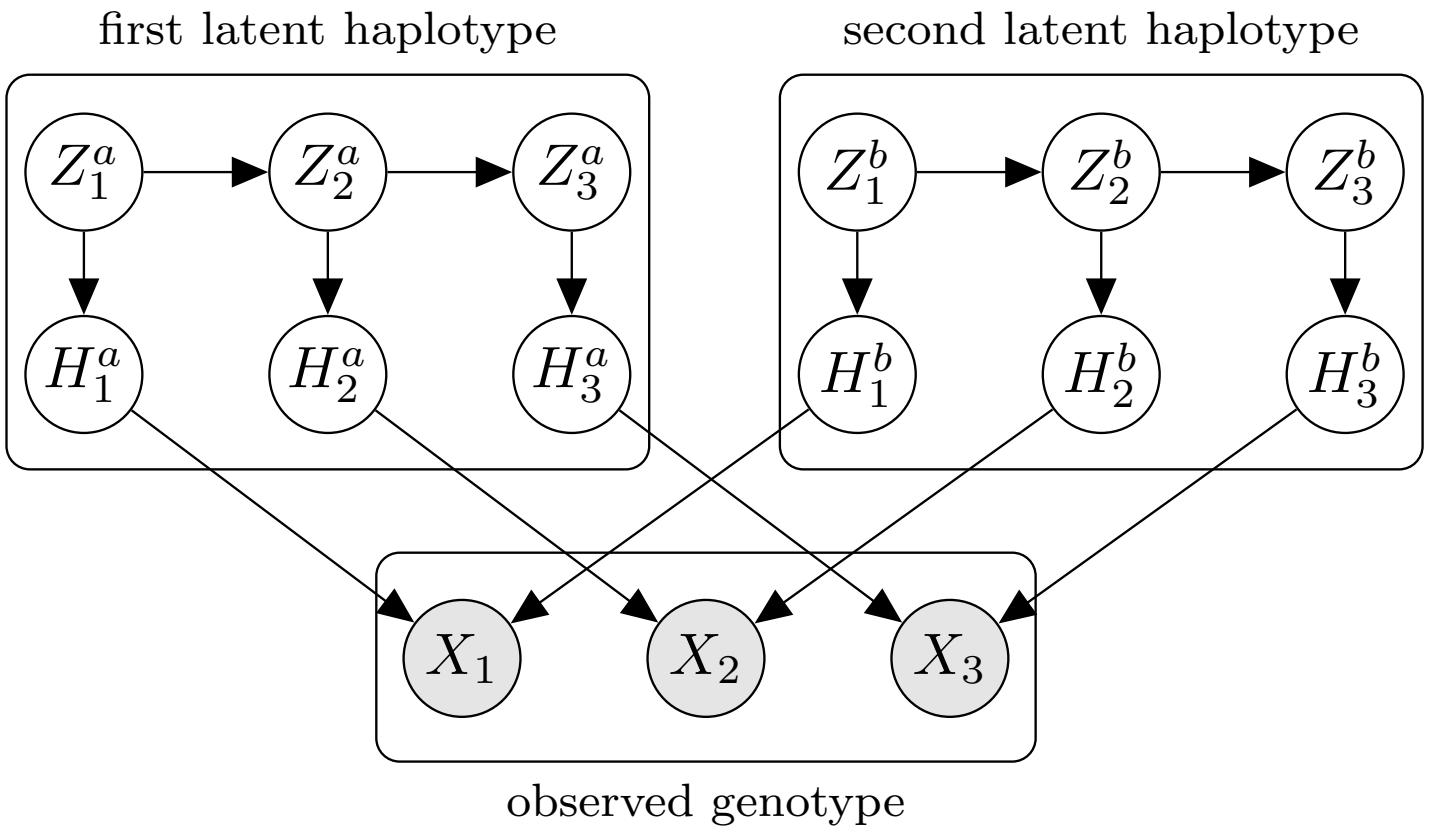
(1) *Sample $\boldsymbol{Z}$ from $p(\boldsymbol{Z}|\boldsymbol{X})$ using forward-backward algorithm*

(2) *Generate a knockoff $\tilde{\boldsymbol{Z}}$ of $\boldsymbol{Z}$ using the SCIP algorithm for a Markov chain*

(3) *Sample $\tilde{\boldsymbol{X}}$ from the emission distribution of $\boldsymbol{X}$ given $\boldsymbol{Z} = \tilde{\boldsymbol{Z}}$*



Knockoffs for genotypes



Experience with data

Crohn's disease (CD)

- Wellcome Trust Case Control Consortium (WTCCC)
- $n \approx 5,000$ subjects ($\approx 2,000$ patients, $\approx 3,000$ healthy controls)
- $p \approx 400,000$ SNPs
- Previously analyzed in WTCCC (2007)

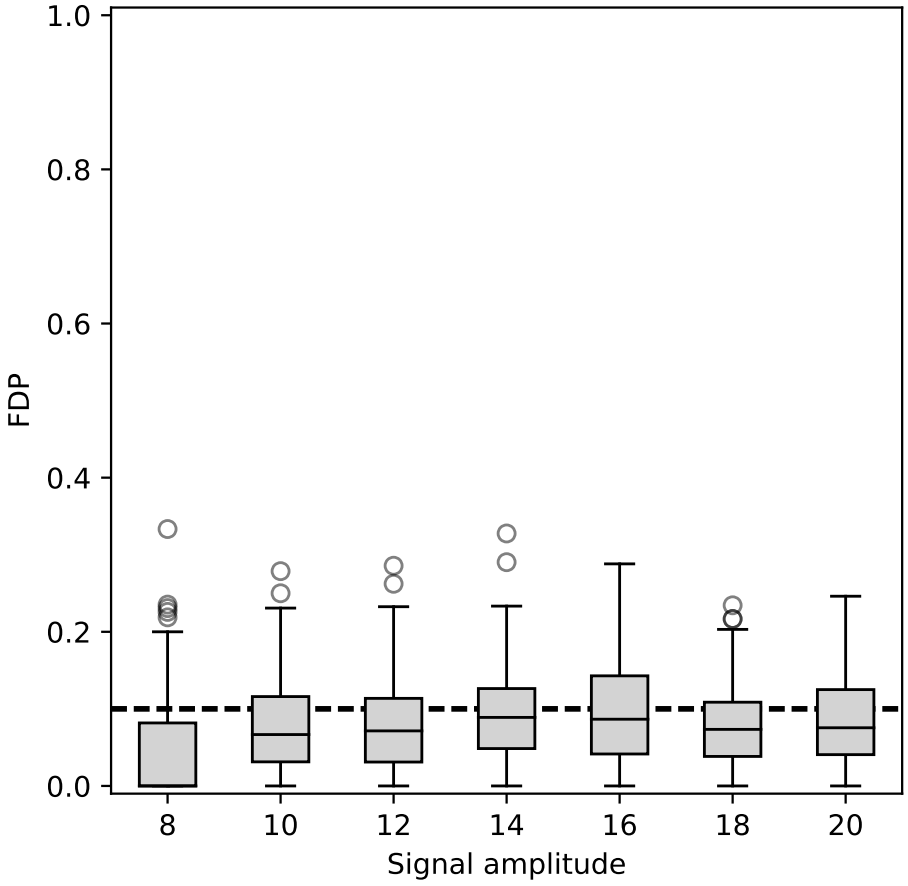
Lipid traits (HDL, LDL cholesterol)

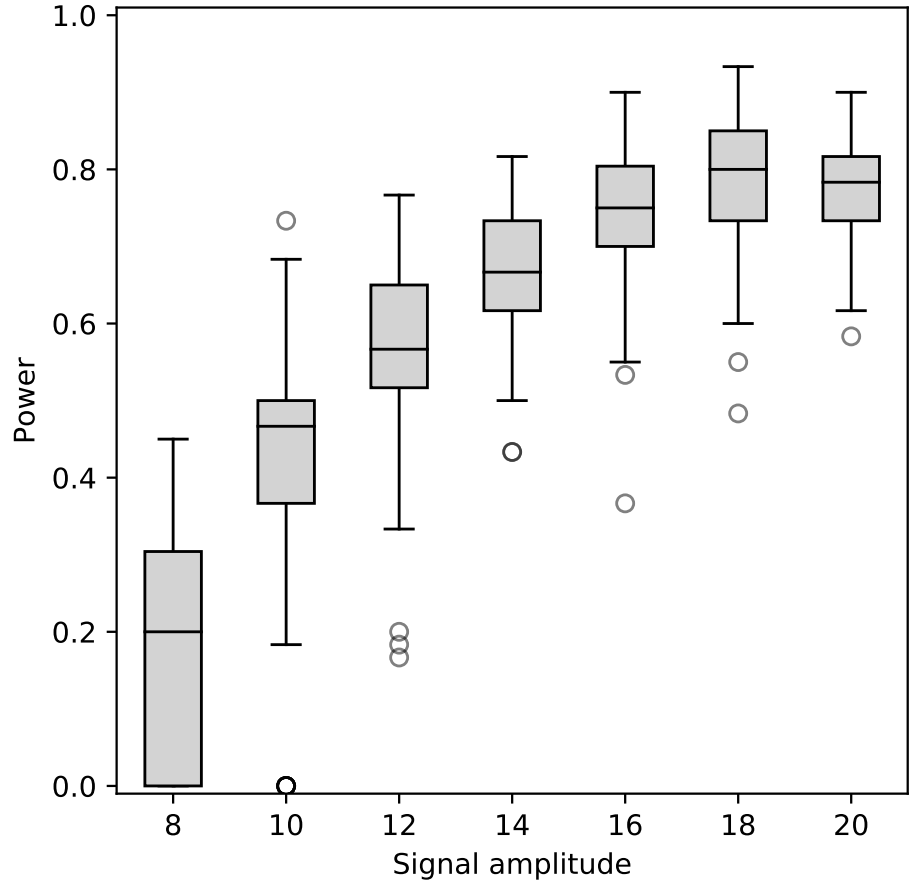
- Northern Finland 1996 Birth Cohort study of metabolic syndrome (NFBC)
- $n \approx 4,700$ subjects
- $p \approx 330,000$ SNPs
- Previously analyzed in Sabatti et al. (2009)

Simulations

- Start from the actual genotypes of 29,258 polymorphisms on chromosome one, genotyped in 14,708 individuals from WTCCC (2007).
- We simulate the response from a conditional logistic regression model of $Y|X$ with 60 non-zero coefficients.
- We prune X to 5260 variables, to guarantee that there is no correlation larger than 0.5. Each variable represents a "cluster" of SNPs
- We fit the fastPHASE model to this data
- Once the parameter are estimated, we construct knockoff copies using the fitted model
- We use logistic regression with ℓ_1-norm penalty tuned by cross- validation.
- Apply knockoff filter at level $q = 0.1$
- We use "clusters" do define false and true discoveries

Simulation results





HMM might not be the "real" distribution of haplotypes, but it works pretty well within this framework (as in the case of imputation).

Application to real data

- Use the same analysis pipeline
- Knockoffs are random: multiple realizations results in different outcomes
- We repeat the procedure multiple times to assess variability
- Compare findings with more recent meta-analysis (Franke et al. 2010, Willer et al. 2013)

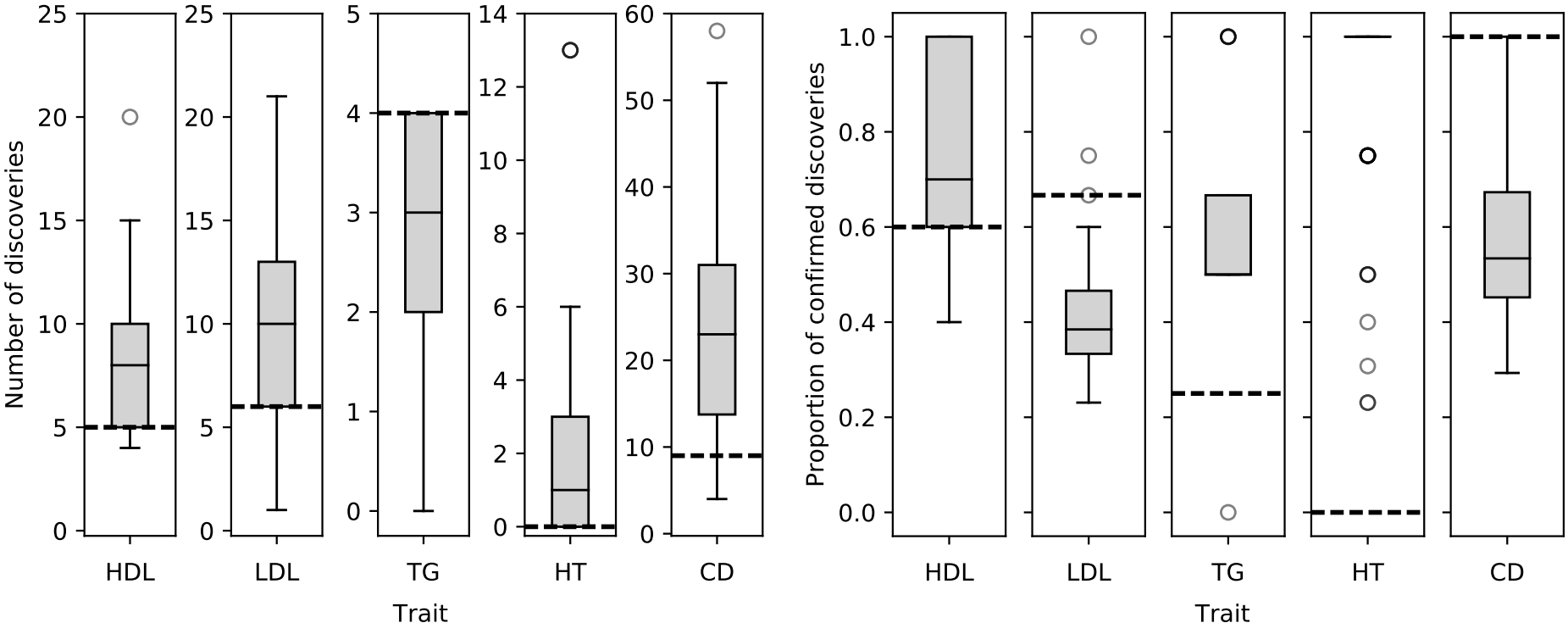
Selection frequency	SNP (cluster size)	Chr.	Position range (Mb)	Confirmed in Willer et al. '13	Found in Sabatti et al. '09
100%	rs1532085 (4)	15	58.68–58.7	yes	yes
100%	rs7499892 (1)	16	57.01–57.01	yes	yes
100%	rs1800961 (1)	20	43.04–43.04	yes	
99%	rs1532624 (2)	16	56.99–57.01	yes	yes
95%	rs255049 (142)	16	66.41–69.41	yes	yes

Table: SNP clusters found to be important for HDL over 100 repetitions of knockoffs.

Selection frequency	SNP (cluster size)	Chr.	Position range (Mb)	Confirmed in Willer et al. '13	Found in Sabatti et al. '09
99%	rs4844614 (34)	1	207.3–207.88		yes
97%	rs646776 (5)	1	109.8–109.82	yes	yes
97%	rs2228671 (2)	19	11.2–11.21	yes	yes
94%	rs157580 (4)	19	45.4–45.41	yes	yes
92%	rs557435 (21)	1	55.52–55.72	yes	
80%	rs10198175 (1)	2	21.13–21.13	yes	yes
76%	rs10953541 (58)	7	106.48–107.3		
62%	rs6575501 (1)	14	95.64–95.64		

Table: SNP clusters found to be important for LDL over 100 repetitions of knockoffs.

Application to real data – overview



Conclusions

- The framework of Model-X knockoff seems appropriate for GWAS
- We can leverage the imputation literature to create working knockoffs
- There is the potential of power gains as sample size increases
- Can these counterfeit genotypes play a role in preserving privacy ?