Biometrika (2018), **xx**, x, *pp*. 1–19 doi: 10.1093/biomet/asm023
Printed in Great Britain

Supplementary Material for Gene Hunting with Hidden Markov Model Knockoffs

BY M. SESIA, C. SABATTI, E. J. CANDÈS
Department of Statistics, Stanford University, Stanford, CA 94305, U.S.A.
msesia@stanford.edu sabatti@stanford.edu candes@stanford.edu

S1. MORE DETAILS ON THE METHODOLODY OF KNOCKOFFS

In § 2·2, we have briefly summarized the knockoff methodology of Candès et al. (2018), which we now outline with further details. Since we have already precisely stated the variable selection problem in § 2·1 and defined the concept of knockoff copies in § 2·2, we now assume that $\tilde{X}$ has been created, and describe how the knockoff methodology proceeds to select relevant variables with provable control of the false discovery rate. As mentioned in § 2·2, the first step is to compute measures of variable importance.

Many options are available for computing the vectors of feature importance statistics T and $\tilde{T}$. For $j = 1, \dots, p$, we require T_j and $\tilde{T}_j$ to be functions of $X, \tilde{X}$ and Y that measure the importance of X_j and $\tilde{X}_j$, respectively, while treating the original variables and the knockoff copies fairly. That is, if we were to swap any subset $S \subseteq \{1, \dots, p\}$ of the original variables with their knockoff copies, this operation should have the only effect of swapping the corresponding elements of T with $\tilde{T}$. Intuitively, this prevents one from cheating by discriminating variables based on whether they are originals or knockoff copies. An example of valid measures of variable importance is

$$T_j = \hat{\beta}_j\Big([X, \tilde{X}], Y; \lambda_{\text{cv}}\Big), \qquad \tilde{T}_j = \hat{\beta}_{j+p}\Big((X, \tilde{X}), Y; \lambda_{\text{cv}}\Big),$$

where term on the left indicates the lasso coefficient for the jth variable obtained by regressing Y on the $2p$ variables in $(X, \tilde{X})$, with regularization parameter λ_{cv} tuned by cross-validation, while the term on the right is the corresponding quantity for the jth knockoff copy. However, this is just a simple example from a multitude of potentially more powerful alternatives. For instance, one could train a random forest and compute T_j and $\tilde{T}_j$ from the gini impurity for the corresponding variables, or develop some other custom statistic specifically tailored for the problem at hand.

The estimated importance measures of the original variables are then compared to those of their corresponding knockoff copies by computing statistics $W_j = w_j(T_j, \tilde{T}_j)$, for some anti-symmetric function w_j, such that $w_j(T_j, \tilde{T}_j) = -w_j(\tilde{T}_j, T_j)$. A typical choice is $W_j = |T_j| - |\tilde{T}_j|$. Properties (1) and (2) from § 2·2 can then be shown to imply that the W_j for all null variables satisfy the property below.

LEMMA S1 (LEMMA 3·3 IN CANDÈS ET AL. (2018)). *Conditional on* $(|W_1|, \dots, |W_p|)$*, the signs of the null* W_j*'s, for* $j \in \mathcal{H}_0$*, are independent and identically distributed coin flips.*

Lemma S1 suggests that the elements of W can be ordered by magnitude and their signs will still effectively be one-bit p-values for the null hypothesis of the corresponding original variable being null. In practice, this means that we can apply the knockoff filter of Barber & Candès (2015) to compute an adaptive significance threshold and select a subset of variables $\hat{S} \subseteq \{1, \dots, p\}$ such that for each $j \in \hat{S}$, W_j is positive and large. The following result from Candès et al. (2018) establishes that this controls the false discovery rate.

THEOREM S1 (THEOREM 3·4 IN CANDÈS ET AL. (2018)). *For some* $\alpha \in (0, 1)$ *and* $c \in \{0, 1\}$*, let the threshold* $\tau > 0$ *of the knockoff filter be defined as*

$$\tau = \min\left\{t > 0 \,:\, \frac{\#\{j \,:\, W_j \le -t\} + c}{\#\{j \,:\, W_j \ge t\}} \le \alpha\right\},$$

and $\tau = +\infty$ if the set is empty. Then the procedure selecting the variables

$$\hat{S} = \{j \,:\, W_j > \tau\}$$

controls at level α the false discovery rate:

$$E\left(\frac{|\hat{S} \cap \mathcal{H}_0|}{|\hat{S}| \vee 1}\right) \leq \alpha,$$

if the offset constant $c = 1$, and the modified false discovery rate:

$$E\left(\frac{|\hat{S} \cap \mathcal{H}_0|}{|\hat{S}| + 1/\alpha}\right) \leq \alpha,$$

if the offset constant $c = 0$.

The results of Theorem S1 are non-asymptotic and hold no matter the dependence between the response and the covariates: the only assumption is that the knockoff copies are constructed for the true population distribution F_X of the covariates.

S2. KNOCKOFFS FOR DISCRETE MARKOV CHAINS

Proof of Proposition 2. By Proposition 1, it suffices to show that Algorithm 1samples $\tilde{X}_j$ from the conditional distribution of X_j given the other original variables X_{-j} and all the knockoff copies $\tilde{X}_{1:(j-1)}$ that have already been sampled: $\tilde{X}_j \sim p(X_j | X_{-j}, \tilde{X}_{1:(j-1)})$, for all $j = 1, \dots, p$.

We proceed by induction, assuming the induction hypothesis that, for some fixed $j \in \{1, \dots, p-1\}$, Algorithm 1samples all knockoff copies $\tilde{X}_i$, for $i \leq j$, from $p(X_i | X_{-i}, \tilde{X}_{1:(i-1)})$, respectively. The main step is to show that $\tilde{X}_{j+1}$ is sampled from $p(X_{j+1} | X_{-(j+1)}, \tilde{X}_{1:j})$.

Let us define $Q_p(k \mid l) = 1$ for all $k, l \in \{1, \dots, K\}$. From the basic properties of conditional probabilities,

$$\begin{aligned}
&\mathbb{P}\big(X_{j+1} = \tilde{x}_{j+1} \mid X_{-(j+1)} = x_{-(j+1)}, \tilde{X}_{1:j} = \tilde{x}_{1:j}\big) \\
&\quad \propto \mathbb{P}\big(X_{j+1} = \tilde{x}_{j+1}, X_{-(j+1)} = x_{-(j+1)}, \tilde{X}_{1:j} = \tilde{x}_{1:j}\big) \\
&\quad \propto \mathbb{P}\big(X_{j+1} = \tilde{x}_{j+1}, X_{-(j+1)} = x_{-(j+1)}, \tilde{X}_{1:(j-1)} = \tilde{x}_{1:(j-1)}\big) \\
&\qquad \times \mathbb{P}\big(\tilde{X}_j = \tilde{x}_j \mid X_{j+1} = \tilde{x}_{j+1}, X_{-(j+1)} = x_{-(j+1)}, \tilde{X}_{1:(j-1)} = \tilde{x}_{1:(j-1)}\big) \\
&\quad \propto \mathbb{P}\big(X_{j+1} = \tilde{x}_{j+1}, X_{-(j+1)} = x_{-(j+1)}\big)\mathbb{P}\big(\tilde{X}_{1:(j-1)} = \tilde{x}_{1:(j-1)} \mid X_{j+1} = \tilde{x}_{j+1}, X_{-(j+1)} = x_{-(j+1)}\big) \\
&\qquad \times \mathbb{P}\big(\tilde{X}_j = \tilde{x}_j \mid X_{j+1} = \tilde{x}_{j+1}, X_{-(j+1)} = x_{-(j+1)}, \tilde{X}_{1:(j-1)} = \tilde{x}_{1:(j-1)}\big).
\end{aligned}$$

Since we are only interested in the dependence on $\tilde{x}_{j+1}$, the first term above can be simplified as:

$$\mathbb{P}\big(X_{j+1} = \tilde{x}_{j+1}, X_{-(j+1)} = x_{-(j+1)}\big) \propto Q_{j+1}(\tilde{x}_{j+1} \mid x_j)\, Q_{j+2}(x_{j+2} \mid \tilde{x}_{j+1}).$$

From the induction hypothesis it follows that the second term is constant with respect to $\tilde{x}_{j+1}$. This is the case because, according to (4), the distribution of $\tilde{X}_i$ only depends on X_{i-1}, X_{i+1} and $\tilde{X}_{i-1}$, for all $i \leq j$. Therefore, the conditional distribution of all $\tilde{X}_{1:(j-1)}$ only depends on $X_{1:j}$.

At this point, we can focus on the third term:

$$\begin{aligned}
&\mathbb{P}\big(\tilde{X}_j = \tilde{x}_j \mid X_{j+1} = \tilde{x}_{j+1}, X_{-(j+1)} = x_{-(j+1)}, \tilde{X}_{1:(j-1)} = \tilde{x}_{1:(j-1)}\big) \\
&\qquad = \frac{Q_j(\tilde{x}_j \mid x_{j-1})\, Q_j(\tilde{x}_j \mid \tilde{x}_{j-1})\, Q_{j+1}(\tilde{x}_{j+1} \mid \tilde{x}_j)}{\mathcal{N}_{j-1}(\tilde{x}_j)\, \mathcal{N}_j(\tilde{x}_{j+1})} \propto \frac{Q_{j+1}(\tilde{x}_{j+1} \mid \tilde{x}_j)}{\mathcal{N}_j(\tilde{x}_{j+1})}.
\end{aligned}$$

The equality above follows from the fact that Algorithm 1samples $\tilde{X}_j$ conditionally independent of X_j, as clearly visible in (4). Thus we can conclude that

$$\mathbb{P}\big(X_{j+1} = \tilde{x}_{j+1} \mid X_{-(j+1)} = x_{-(j+1)}, \tilde{X}_{1:j} = \tilde{x}_{1:j}\big) \propto Q_{j+1}(\tilde{x}_{j+1} \mid x_j)\, Q_{j+2}(x_{j+2} \mid \tilde{x}_{j+1}) \frac{Q_{j+1}(\tilde{x}_{j+1} \mid \tilde{x}_j)}{\mathcal{N}_j(\tilde{x}_{j+1})}.$$

This proves that the induction hypothesis also holds for $j+1$. The special case $j = 1$ remains to be considered. However, this is straightforward since Algorithm 1samples $\tilde{X}_1$, independently of X_1, from

$$\mathbb{P}\big(X_1 = \tilde{x}_1 \mid X_{-1} = x_{-1}\big) = \mathbb{P}\big(X_1 = \tilde{x}_1 \mid X_2 = x_2\big) \propto \mathbb{P}\big(X_1 = \tilde{x}_1, X_2 = x_2\big) = q_1(\tilde{x}_1) Q_2(x_2 \mid \tilde{x}_1).$$

S3. Sampling latent paths for a hidden Markov model

Algorithm 2for generating a knockoff copy of a hidden Markov model requires sampling from the conditional distribution of the latent variables Z, given all the observable variables X. This task is closely related to that of finding the most likely a-posteriori sequence of hidden states, i.e. the Viterbi path, and it can be solved efficiently with forward-backward sampling as in Algorithm 3. Earlier examples of this technique are found in Zhu et al. (1998), in the context of biological sequence alignment. A similar method is also described in Cawley & Pachter (2003), where instead of proceeding as we suggest, they first compute a collection of backward probabilities and then sample Z with a forward pass. For completeness, we prove here the correctness of Algorithm 3.

Proof of Proposition 3. For each variable $j \in \{1, \dots, p\}$, we define the forward probability

$$\alpha_j(k) = \mathbb{P}\big(x_{1:j}, Z_j = k\big),$$

which is the probability of observing the features $X_{1:j} = x_{1:j}$ up to time j and ending up in the hidden state k. Note that for $j = 1$ this is simply

$$\alpha_1(k) = q_1(k)\, f_1(x_1 \mid k),$$

where $q_1(k)$ is the marginal distribution of Z_1. The other forward probabilities can be computed recursively as follows:

$$\begin{aligned}
\alpha_{j+1}(k) = \mathbb{P}\big(x_{1:(j+1)}, Z_{j+1} = k\big) &= \sum_l \mathbb{P}\big(x_{j+1}, Z_{j+1} = k \mid Z_j = l, x_{1:j}\big) \alpha_j(l) \\
&= \sum_l \mathbb{P}\big(x_{j+1} \mid Z_{j+1} = k\big)\, \mathbb{P}\big(Z_{j+1} = k \mid Z_j = l\big) \alpha_j(l) \\
&= f_{j+1}(x_{j+1} \mid k) \sum_l Q_{j+1}(k \mid l)\, \alpha_j(l).
\end{aligned}$$

These equations can be written more compactly in matrix notation:

$$\alpha_j = (Q_j \alpha_{j-1}) \odot \beta_j, \qquad \beta_j(k) = f_j(x_j \mid k),$$

where $\odot$ indicates component-wise multiplication.

Having computed the forward probabilities in the forward pass, we can now sample from $p(Z \mid X)$, starting from Z_p and back-tracking along the sequence all the way to Z_1. Our approach arises naturally from:

$$\mathbb{P}\big(Z_{1:p} = z_{1:p} \mid x_{1:p}\big) = \mathbb{P}\big(Z_{1:(p-1)} = z_{1:(p-1)} \mid Z_p = z_p, x_{1:p}\big)\, \mathbb{P}\big(Z_p = z_p \mid x_{1:p}\big).$$

This identity suggests that one should start by sampling z_p from the discrete distribution

$$\mathbb{P}\big(Z_p = z_p \mid x_{1:p}\big) = \frac{\alpha_p(z_p)}{\sum_k \alpha_p(k)}.$$

Once z_p is chosen, we can think of it as a fixed parameter and turn on to sampling the random variable Z_{p-1}. To this end, note that

$$\begin{aligned}\mathbb{P}\big(Z_{1:(p-1)} = z_{1:(p-1)} \mid z_p, x_{1:p}\big) &= \mathbb{P}\big(Z_{1:(p-1)} = z_{1:(p-1)} \mid z_p, x_{1:p-1}\big) \\ &= \mathbb{P}\big(Z_{1:(p-2)} = z_{1:(p-2)} \mid Z_{p-1} = z_{p-1}, x_{1:p}\big) \underbrace{\mathbb{P}\big(Z_{p-1} = z_{p-1} \mid z_p, x_{1:(p-1)}\big)}_{\propto Q_p(z_p|z_{p-1})\,\alpha_{p-1}(z_{p-1})}.\end{aligned}$$

Hence, we sample z_{p-1} from

$$\mathbb{P}\big(Z_{p-1} = z_{p-1} \mid z_p, x_{1:(p-1)}\big) = \frac{Q_p(z_p \mid z_{p-1})\,\alpha_{p-1}(z_{p-1})}{\sum_k Q_p(z_p \mid k)\,\alpha_{p-1}(k)}.$$

We continue in this fashion and, at step $p - j + 1$, we sample z_j from

$$\mathbb{P}\big(Z_j = z_j \mid z_{j+1}, x_{1:j}\big) = \frac{Q_{j+1}(z_{j+1} \mid z_j)\,\alpha_j(z_j)}{\sum_k Q_{j+1}(z_{j+1} \mid k)\,\alpha_j(k)}.$$

This completes the proof.

To summarize, in the first phase of Algorithm 3the forward variables are computed with the forward pass. Then, sampling is done with a backward pass. This process allows one to sample a complete path of latent variables from their conditional law given the corresponding emitted variables X. Since this algorithm only involves matrix multiplications and other trivial operations, its computation time is $O(pK^2)$, where K is the size of the state space of the latent Markov chain. This complexity is the same as that of Algorithm 1for generating knockoff copies of a Markov chain.

S4. ADDITIONAL DETAILS FOR THE NUMERICAL SIMULATION WITH REAL GENETIC COVARIATES

The setup for our numerical simulation with real genetic covariates follows the footsteps of Candès et al. (2018), but we repeat the details here for completeness.

As covariates, we use 29,258 polymorphisms on chromosome one, from 14,708 individuals genotyped by the Wellcome Trust Case Control Consortium (WTCCC, 2007). These were obtained by combining the genetic information from the healthy subjects and the Crohn's disease patients with that from 5 other diseases from the same data set: coronary artery disease, hypertension, rheumatoid arthritis, type-1 diabetes and type-2 diabetes. Conditional on $X = (X_1, \ldots, X_p)$, the response Y is sampled from a binomial generalized linear model with a logit link function. The coefficient vector β has 60 non-zero elements, chosen uniformly at random, which correspond to the set $\mathcal{S}$ of relevant features. The signs the non-zero coefficients in β are independent coin flips. In summary,

$$Y \mid X \sim \text{Bernoulli}\left(\text{logit}\left(X^T\beta\right)\right), \qquad \text{where} \quad \beta_j = \begin{cases} \frac{a\,s_j}{\sqrt{n}}, & j \in \mathcal{S}, \\ 0, & \text{otherwise.} \end{cases}$$

Above, the signal amplitude a is a parameter that we can vary in the simulations, while s_j indicates the sign of β_j. The signs are chosen independently such that $s_j = +1$ with probability $0{\cdot}5$ and $s_j = -1$ otherwise. Furthermore, since variables corresponding to different polymorphisms have very different marginal distributions, we standardize the covariates X so that they have mean 0 and variance 1, before sampling from the

conditional logistic model described above. This allows us to reduce the variability of our experiments due to the random choice of the model.

Once the response vector Y is sampled, we prune the variables to remove extremely high correlations, as motivated in § S7·2. First, we hold out a random subset of 1000 observations. Then, we split the remaining samples, i.e. the rows of X, into 10 subsets with approximately 1400 observations each. For each subset, we separately apply hierarchical clustering of the empirical covariance matrix, as described in § S7·2, and use the common 1000 hold-out observations for choosing the cluster representatives, instead of further splitting the data. This somewhat involved procedure, inspired by Candès et al. (2018), allows us to obtain 10 datasets with real covariates and a simulated response, such that in each of them the number of samples is comparable to the data analysis in § 7.

Once pruning is performed, for each of the 10 splits we separately fit the model of § 5 with fastPHASE, sample the knockoff copies and perform variable selection with our procedure on each of the 10 data subsets, as described in § 6. As mentioned in § S7·2, the observations used for selecting the cluster representatives can be partially re-used to fit F_X and compute the feature importance measures each time, making sure that their corresponding knockoff copies are set identical to the originals in order to avoid any selection bias. Since at this point we only have 10 point estimates for the power and the false discovery rate, we repeat the entire procedure 10 times, starting from the choice of the logistic model and re-sampling the response Y. Thus, in the end we obtain 100 point estimates in the unconditional model, with 10 random samples of X and 10 samples of $Y \mid X$ for each one of them.

S5. Numerical simulation with Markov chains

S5·1. *A toy model*

We consider a vector X of $p = 1000$ covariates distributed as a discrete Markov chain taking values in a state space $\mathcal{X} = \{-2, -1, 0, +1, +2\}$ of size $K = |\mathcal{X}| = 5$. In the notation of (3), this can be written as $X \sim \text{MC}(q_1, Q)$, with an initial distribution q_1 assumed to be uniform on $\mathcal{X}$. For each $j \in \{1, \dots, p-1\}$, we set:

$$Q_j(k \mid l) = \begin{cases} \frac{1}{K} + \gamma_j \left(1 - \frac{1}{K}\right), & k = l, \\ \left[1 - \frac{1}{K} - \gamma_j \left(1 - \frac{1}{K}\right)\right] \frac{1}{K-1}, & k \neq l, \end{cases}$$

where the hyper-parameters γ_j are once randomly sampled γ_j independently from the uniform distribution on $[0, 0{\cdot}5]$ and then held constant.

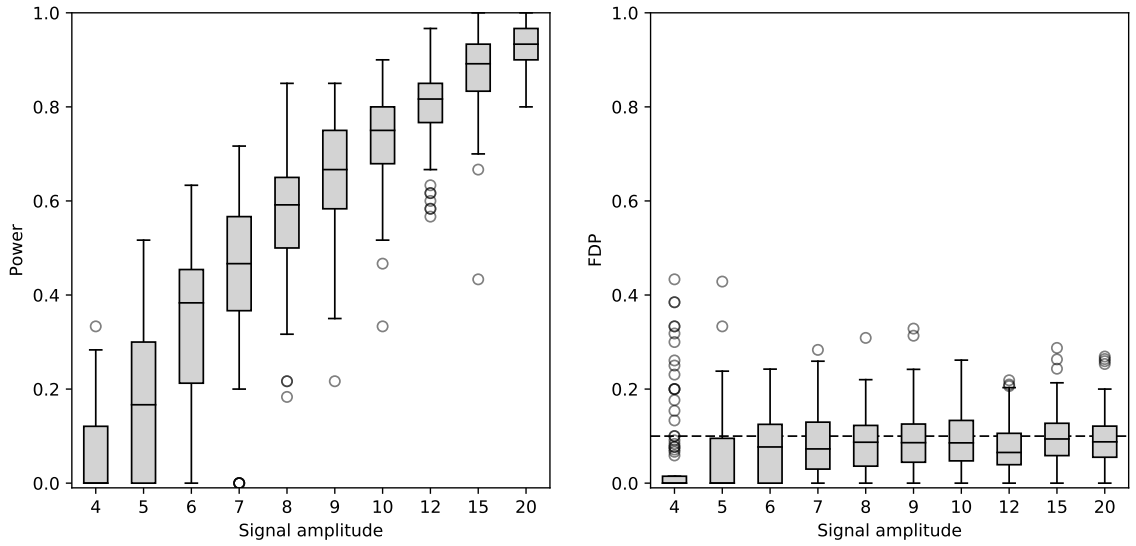
Conditional on $X = (X_1, \dots, X_p)$, the response Y is sampled from a binomial generalized linear model with a logit link function. The coefficient vector β has 60 non-zero elements, which correspond to the set $\mathcal{S}$ of relevant features. In summary,

$$Y \mid X \sim \text{Bernoulli}\left(\text{logit}\left(X^T \beta\right)\right), \qquad \text{where} \quad \beta_j = \begin{cases} \frac{a}{\sqrt{n}}, & j \in \mathcal{S}, \\ 0, & \text{otherwise.} \end{cases}$$

Above, the signal amplitude a is a parameter that we can vary in the simulations.

S5·2. *Effect of signal amplitude*

We draw 1000 independent observations of (X, Y) from the model described above. For different values of the signal amplitude a, we apply the knockoff construction procedure for Markov chains, using the true model parameters (q_1, Q). It is interesting to note that, since $p = n$, the observations are perfectly separable, i.e. there exists a hyperplane in the feature space that perfectly separates the two classes of Y, and the maximum



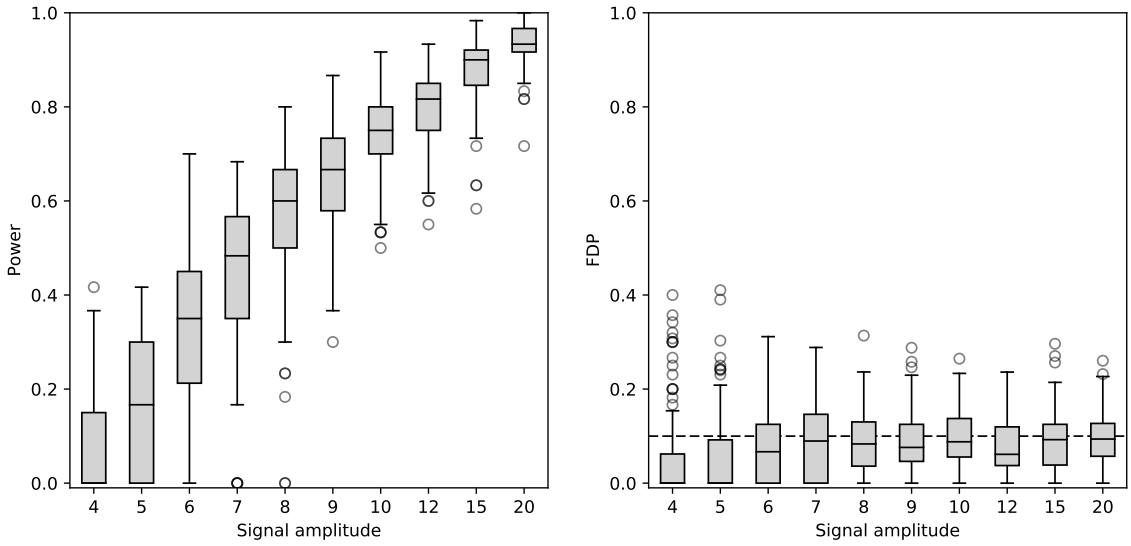
(a) (b)

Fig. S1: Power (a) and false discovery proportion (b) of our procedure in a simulation with $n = 1000$ and $p = 1000$, over 100 independent experiments. Variables are distributed as a discrete Markov chain. The knockoff copies are constructed using the true model parameters. The response $Y \mid X$ is sampled from a logistic regression model. The dashed black line in (b) indicates the target level $\alpha = 0{\cdot}1$.

likelihood estimate of β, therefore, does not exist. This is the reason why it is useful to leverage some sparsity in order to identify the relevant variables. As variable importance measures, we compute $W_j = |\hat{\beta}_j(\lambda_{\text{cv}})| - |\hat{\beta}_{j+p}(\lambda_{\text{cv}})|$, where $\hat{\beta}_j(\lambda_{\text{cv}})$ and $\hat{\beta}_{j+p}(\lambda_{\text{cv}})$ are the logistic regression coefficients for the jth variable and its knockoff copy, respectively, regularized with an ℓ_1-norm penalty chosen by 10-fold cross-validation. Finally, we estimate the set of relevant variables using the knockoff filter with offset $c = 1$ and target level $\alpha = 0{\cdot}1$. The results shown in Figure S1 and Table S1 correspond to 100 independent replications of this experiment. Empirically, our method is confirmed to control the false discovery rate for all values of the signal amplitude. As it should be expected, the actual false discovery proportion is not always below the target value, but is quite concentrated around its mean.

S5·3. *Robustness to overfitting*

In the previous example, we generated the knockoff variables using the real distribution of X. However, in most practical applications this is not known exactly and it must be estimated from the available data. In a more realistic situation one may have some prior knowledge that a Markov chain is a good model for the covariates, but ignore the exact form of the transition matrices. Therefore, we repeat the previous experiment, generating instead the knockoff copies $\tilde{X}$ from the fitted values of the Markov chain parameters. The estimates $(\hat{q}_1, \hat{Q})$ are obtained by maximum-likelihood with Laplace smoothing on all the available observations of X. This is a well-known technique that can be used to improve the estimation of the transition matrices. In order to avoid estimating any transition probabilities as zero, we simply add one to all transition counts. The results shown in Figure S2 and Table S1 are very similar to those of Figure S1. This shows that the false discovery rate is still controlled, and it also suggests that our procedure is robust to fitting the feature distribution.



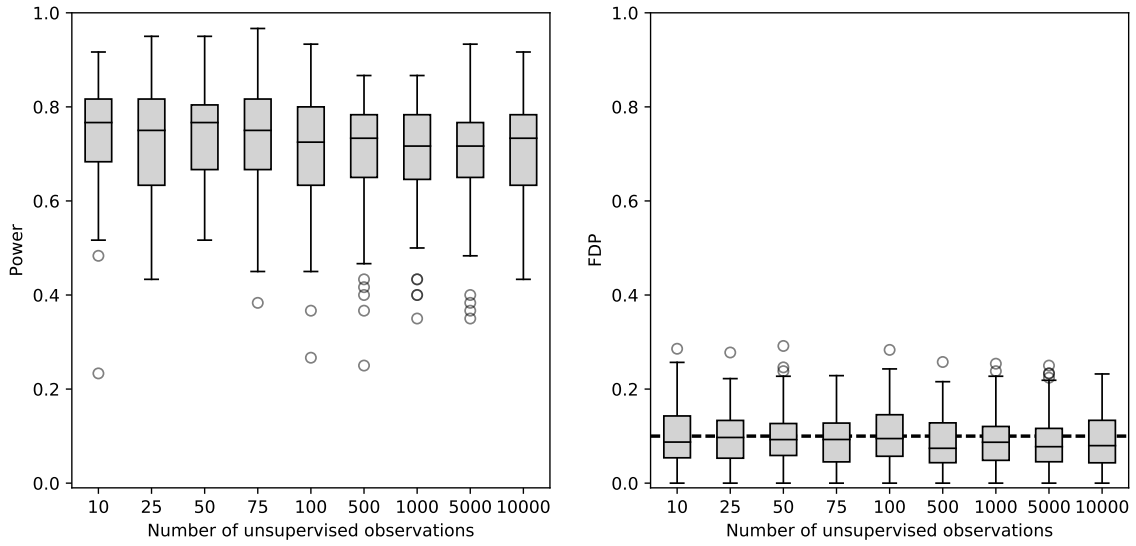
(a) (b)

Fig. S2: Power (a) and false discovery proportion (b) of our procedure with simulated Markov chain covariates, with knockoffs sampled using estimates of the transition matrices obtained from the same dataset. The setup is otherwise the same as that in Figure S1.

Signal	True F_X		Estimated F_X	
amplitude	FDR (95% c.i.)	Power (95% c.i.)	FDR (95% c.i.)	Power (95% c.i.)
4	$0{\cdot}050 \pm 0{\cdot}020$	$0{\cdot}051 \pm 0{\cdot}018$	$0{\cdot}054 \pm 0{\cdot}020$	$0{\cdot}064 \pm 0{\cdot}020$
5	$0{\cdot}057 \pm 0{\cdot}017$	$0{\cdot}154 \pm 0{\cdot}031$	$0{\cdot}062 \pm 0{\cdot}019$	$0{\cdot}155 \pm 0{\cdot}031$
6	$0{\cdot}083 \pm 0{\cdot}014$	$0{\cdot}329 \pm 0{\cdot}034$	$0{\cdot}078 \pm 0{\cdot}015$	$0{\cdot}312 \pm 0{\cdot}035$
7	$0{\cdot}084 \pm 0{\cdot}014$	$0{\cdot}446 \pm 0{\cdot}031$	$0{\cdot}091 \pm 0{\cdot}015$	$0{\cdot}449 \pm 0{\cdot}031$
8	$0{\cdot}086 \pm 0{\cdot}012$	$0{\cdot}566 \pm 0{\cdot}025$	$0{\cdot}089 \pm 0{\cdot}013$	$0{\cdot}560 \pm 0{\cdot}029$
9	$0{\cdot}092 \pm 0{\cdot}013$	$0{\cdot}658 \pm 0{\cdot}024$	$0{\cdot}088 \pm 0{\cdot}013$	$0{\cdot}653 \pm 0{\cdot}023$
10	$0{\cdot}093 \pm 0{\cdot}011$	$0{\cdot}730 \pm 0{\cdot}020$	$0{\cdot}096 \pm 0{\cdot}011$	$0{\cdot}741 \pm 0{\cdot}017$
15	$0{\cdot}096 \pm 0{\cdot}011$	$0{\cdot}874 \pm 0{\cdot}016$	$0{\cdot}092 \pm 0{\cdot}012$	$0{\cdot}878 \pm 0{\cdot}014$
20	$0{\cdot}094 \pm 0{\cdot}011$	$0{\cdot}930 \pm 0{\cdot}009$	$0{\cdot}098 \pm 0{\cdot}011$	$0{\cdot}933 \pm 0{\cdot}009$

Table S1: False discovery rate and average power in the numerical experiments of Figure S1 and S2. We compare the results obtained with knockoff variables created using the exact (left) and estimated (right) Markov chain model parameters.

Alternatively, if additional unsupervised samples are available, one can use them to improve the estimation of the covariate distribution. We illustrate this idea by generating unlabeled datasets of varying size n_{u}, from the same population. In principle, one could use both the supervised and the unsupervised observations of X to estimate the parameters of F_X. However, we choose to fit the parameters only on the latter, in order to better observe the effect of overfitting. For a range of values of n_{u}, we compute $(\hat{q}_1, \hat{Q})$ and proceed as in the previous examples, repeating the experiment 100 times. The results are shown in Figure S3. We observe that our procedure is robust to overfitting. Even in the extreme cases in which n_u is very small, i.e. $n_u \leq 50$, the



(a) (b)

Fig. S3: Power (a) and false discovery proportion (b) of our procedure with simulated Markov chain covariates. Knockoffs are sampled using estimates of the transition matrices obtained from an independent dataset of n_u unsupervised observations of X, for different values of n_u. The signal amplitude is $a = 10$. The setup is otherwise the same as that in Figure S1.

empirical false discovery rate is below the nominal value, while for larger values of n_u the validity of the false discovery rate control is clear.

S6. NUMERICAL SIMULATION WITH HIDDEN MARKOV MODELS

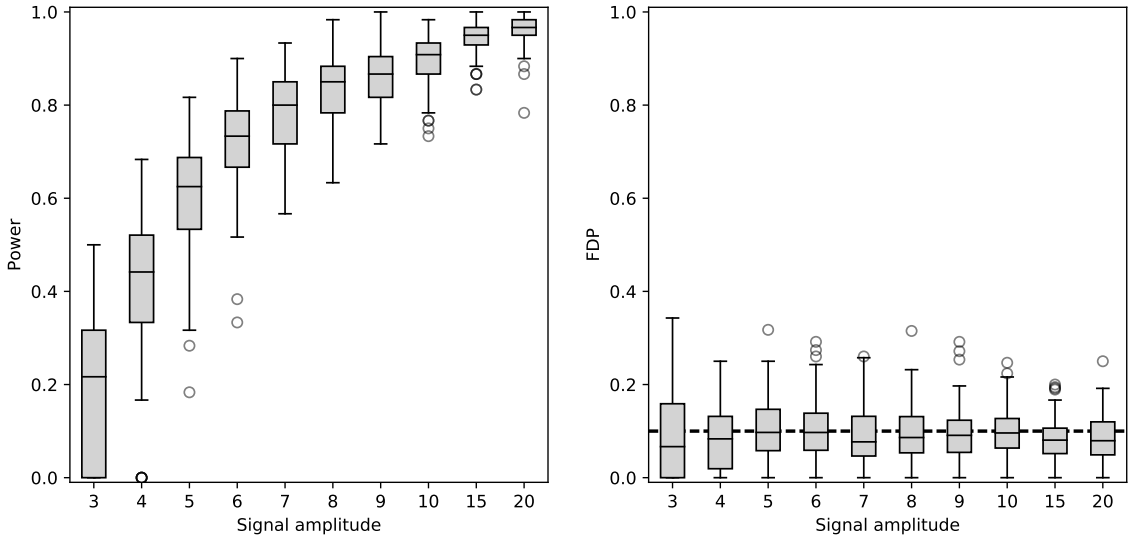
S6·1. *A toy model*

We consider a vector X of 1000 covariates distributed as the hidden Markov model defined below. The parametrization that we adopt is loosely inspired by the left-right models used for speech recognition (Juang & Rabiner, 1991), but we do not aim to realistically simulate any specific application. Instead, we prefer to keep the model extremely simple for the sake of exposition. Here, the latent Markov chain $Z \sim \text{MC}(q_1, Q)$ takes on values in $\{0, 1, \dots, K-1\}^p$ and its states evolve clockwise according to

$$q_1(k) = \begin{cases} 1, & k = 1, \\ 0, & \text{otherwise}, \end{cases} \qquad Q_j(k \mid l) = \begin{cases} 0{\cdot}9, & k = l, \\ 0{\cdot}1, & k = l + 1 \mod K, \\ 0, & \text{otherwise}, \end{cases} \qquad j \in \{2, \dots, p\},$$

for $k, l \in \{0, 1, \dots, K-1\}$. Concretely, we let $K = 9$ and we assume for simplicity that all observed variables X_j take on values in a set $\mathcal{X} = \{-4, -3, \dots, +3, +4\}$, also of size K. The emission probabilities $f_j(x \mid z)$ are defined, for some $\gamma \in (0, 1)$, as

$$f_j(x \mid z) = \begin{cases} \frac{\gamma}{2}, & (x+4) = z \text{ or } (x+4) = z+1, \\ \frac{\gamma}{2}, & (x+4) = 0 \text{ and } z = K-1, \\ \frac{1-\gamma}{K-2}, & \text{otherwise}. \end{cases}$$



(a) (b)

Fig. S4: Power (a) and false discovery proportion (b) of our procedure in a simulation with $n = 1000$ and $p = 1000$, over 100 independent experiments. Variables are distributed as a discrete hidden Markov model. The knockoff copies are constructed using the true model parameters. The response $Y \mid X$ is sampled from a logistic regression model. The dashed black line in (b) indicates the target level $\alpha = 0{\cdot}1$.

In this example, we set $\gamma = 0{\cdot}35$ because we have observed empirically that it yields an interesting structure with moderately strong correlations.

Conditional on $X = (X_1, \ldots, X_p)$, the response Y is sampled from the same binomial generalized linear model of § S5. Again, we vary the signal amplitude in the simulations.

S6·2. *Effect of signal amplitude*

We simulate 1000 independent observations of (X, Y) from the model described above. For different values of the signal amplitude a, we apply our method to construct knockoff copies of the hidden Markov model, using the exact model parameters. We select relevant variables after computing the same importance measures as in § S5, and applying the knockoff filter with offset $c = 1$ and target level $\alpha = 0{\cdot}1$. The power and false discovery proportion shown in Figure S4 and Table S2 correspond to 100 independent replications of this experiment. The results confirms that our procedure accurately controls the false discovery rate for all values of the signal amplitude.

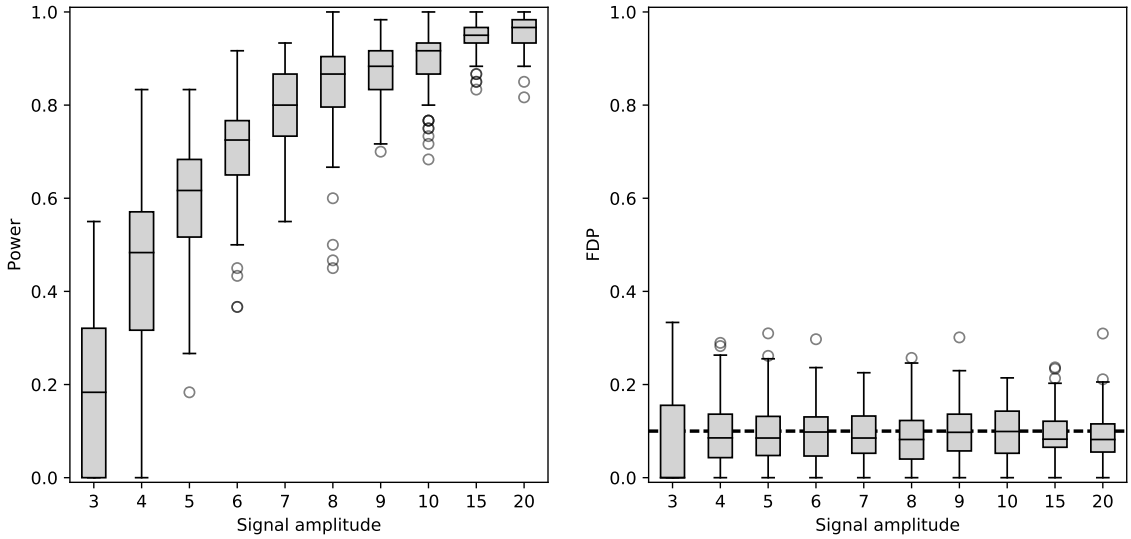
S6·3. *Robustness to overfitting*

In the previous example, we have sampled the knockoff variables by exploiting our knowledge of the true distribution of X. Now, we continue as in § S5 to verify the robustness of our procedure to the estimation of F_X. Instead of using the exact values of (q_1, Q, f), we fit them on the available data using the Baum-Welch algorithm (Rabiner, 1989). The power and false discovery rate shown in Figure S5 and Table S2 are estimated over 100 replications, for different values of the signal amplitude. Similarly to the earlier example with Markov chain covariates, our technique behaves robustly and maintains control as expected.

Signal	True F_X		Estimated F_X	
amplitude	FDR (95% c.i.)	Power (95% c.i.)	FDR (95% c.i.)	Power (95% c.i.)
2	$0{\cdot}037 \pm 0{\cdot}019$	$0{\cdot}030 \pm 0{\cdot}014$	$0{\cdot}049 \pm 0{\cdot}025$	$0{\cdot}029 \pm 0{\cdot}013$
3	$0{\cdot}091 \pm 0{\cdot}019$	$0{\cdot}196 \pm 0{\cdot}028$	$0{\cdot}078 \pm 0{\cdot}019$	$0{\cdot}189 \pm 0{\cdot}033$
4	$0{\cdot}082 \pm 0{\cdot}013$	$0{\cdot}414 \pm 0{\cdot}030$	$0{\cdot}094 \pm 0{\cdot}014$	$0{\cdot}432 \pm 0{\cdot}040$
5	$0{\cdot}102 \pm 0{\cdot}013$	$0{\cdot}610 \pm 0{\cdot}023$	$0{\cdot}094 \pm 0{\cdot}013$	$0{\cdot}592 \pm 0{\cdot}026$
6	$0{\cdot}105 \pm 0{\cdot}012$	$0{\cdot}726 \pm 0{\cdot}020$	$0{\cdot}093 \pm 0{\cdot}011$	$0{\cdot}708 \pm 0{\cdot}022$
7	$0{\cdot}090 \pm 0{\cdot}012$	$0{\cdot}781 \pm 0{\cdot}017$	$0{\cdot}093 \pm 0{\cdot}011$	$0{\cdot}790 \pm 0{\cdot}018$
8	$0{\cdot}093 \pm 0{\cdot}012$	$0{\cdot}830 \pm 0{\cdot}015$	$0{\cdot}086 \pm 0{\cdot}011$	$0{\cdot}839 \pm 0{\cdot}020$
9	$0{\cdot}093 \pm 0{\cdot}011$	$0{\cdot}865 \pm 0{\cdot}013$	$0{\cdot}099 \pm 0{\cdot}010$	$0{\cdot}877 \pm 0{\cdot}012$
10	$0{\cdot}097 \pm 0{\cdot}009$	$0{\cdot}896 \pm 0{\cdot}011$	$0{\cdot}099 \pm 0{\cdot}011$	$0{\cdot}898 \pm 0{\cdot}012$
15	$0{\cdot}083 \pm 0{\cdot}009$	$0{\cdot}945 \pm 0{\cdot}007$	$0{\cdot}093 \pm 0{\cdot}010$	$0{\cdot}950 \pm 0{\cdot}007$
20	$0{\cdot}086 \pm 0{\cdot}009$	$0{\cdot}965 \pm 0{\cdot}006$	$0{\cdot}092 \pm 0{\cdot}010$	$0{\cdot}954 \pm 0{\cdot}007$

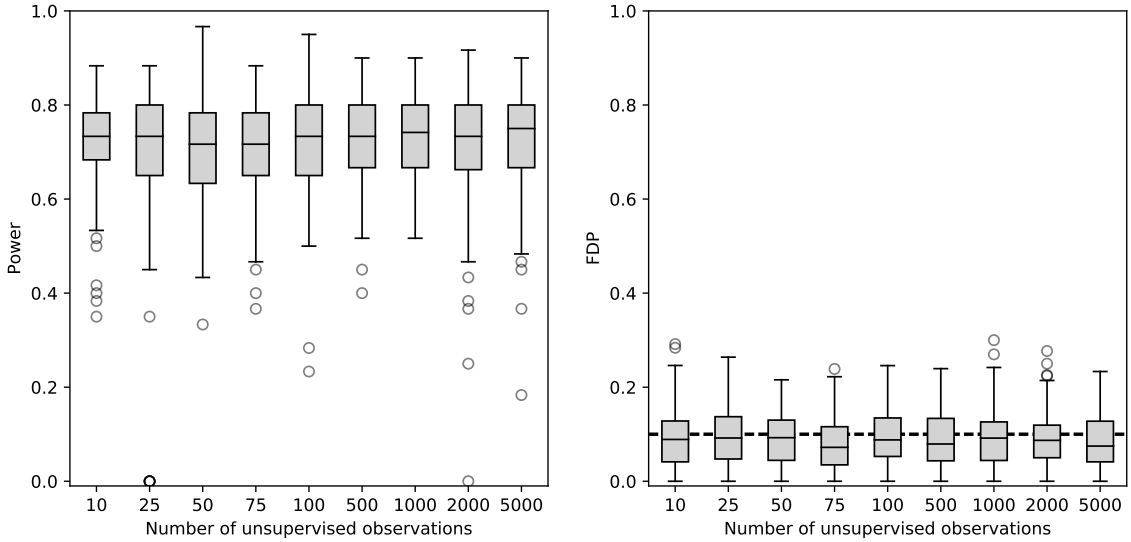
Table S2: False discovery rate and average power in the numerical experiments of Figure S4 and S5. We compare the results obtained with knockoff variables created using the exact (left) and estimated (right) parameters.

Finally, we repeat the experiment by fitting the parameters on an independent and unsupervised dataset of size n_{u}, for different values of n_u. The results are shown in Figure S6 and they correspond to a range of values for n_u and fixed signal amplitude $a = 6$. Again, the false discovery rate is consistently controlled. It should not be suprising that this works even when n_u is as small as 10. Unlike the numerical experiments with the Markov chain variables considered earlier, the transition matrices and emission probabilities for this hidden Markov model are homogeneous for all covariates, i.e. $Q_j = Q_{j+1}$, for all j. This simple model results in fewer parameters to be estimated, thus contributing to the overall robustness.



(a) (b)

Fig. S5: Power (a) and false discovery proportion (b) of our procedure with simulated hidden Markov model covariates, with knockoffs sampled using the parameters fitted with the expectation-maximization algorithm on the same dataset. The setup is otherwise the same as that in Figure S4.



(a) (b)

Fig. S6: Power (a) and false discovery proportion (b) of our procedure with simulated hidden Markov model covariates. Knockoffs are sampled using parameter estimates obtained with the expectation-maximization algorithm from an independent dataset of n_u unlabeled observations of X, for different values of n_u. The signal amplitude is $a = 6$. The setup is otherwise the same as in Figure S4.

S7. ADDITIONAL DETAILS FOR THE DATA ANALYSIS

S7·1. *Data pre-processing*

We follow the pre-processing steps of Sabatti et al. (2009) and Barber & Candès (2016) for the metabolic syndrome data. This reduces the total number of polymorphisms to $328,934$. Cholesterol and triglycerides levels are log-transformed prior to analysis, and all response variables are regressed on the top five principal components of the genotype matrix to correct for population stratification (Price et al., 2006). The residuals from these regressions define the phenotypes we actually analyze. Following the footsteps of earlier literature, we concluded that the Crohn's disease data does not require additional pre-processing (WTCCC, 2007). A summary of both datasets is shown in Table S3.

S7·2. *Variable pruning*

The presence of extremely high correlations between neighboring polymorphisms is a well-known issue in genotype association studies. The traditional approach for dealing with this is to perform marginal tests for each site individually and then group those findings that occurr within the same small physical region. In the end, studies often report regions that contain interesting variants, rather than individual polymorphisms. This approach can be justified within the framework of marginal testing with family-wise error rate control, but it is dangerous when trying to control the false discovery rate. Indeed, a-posteriori aggregation can inflate the false discovery rate because the number of discoveries is artificially increased by the presence of multiple sites belonging to the same highly correlated region, which will be later reported as a whole. This issue has been addressed before in special cases (Pacifico et al., 2004; Siegmund et al., 2011), but the problem remains that a-posteriori aggregation is intrinsically ill-suited for high-dimensional problems. With a limited number of samples, one cannot test hypotheses at very high resolution: in a typical association study we simply do not have enough observations to distinguish between near-identical variables.

In the formulation of the variable selection problem that we adopt, a-priori aggregation of the hypotheses is a natural solution. By pruning the variables and applying the method of knockoff to the representatives identified from each highly-correlated region, we obtain easily intepretable findings and maintain control of the false discovery rate. Our pruning procedure follows along the steps of Candès et al. (2018) and proceeds as follows. First, we apply single-linkage hierarchical clustering to the complete set of variables, using the empirical correlation matrix as a similarity measure, and obtaining a clustering dendrogram. Then, we identify groups by cutting the dendrogram at the lowest possible height such that the highest correlation between any two distinct groups does not exceed a certain threshold. In this paper, we use a threshold equal to $0{\cdot}5$, chosen before looking at the response Y and before selecting any variables. This value was heuristically determined as to obtain clusters that are not too large and that span a range of physical positions that is comparable in size with those typically reported in the literature on genome-wide association studies. Once the groups are found, we choose their representatives. For each group, we choose as its representative the single polymorphism within it that is the most closely associated to the response, based on a marginal t-test performed using a hold-out subset of 20% of the original n samples. Finally, we remark that by pruning we are only reducing the largest correlations among our predictors to $0{\cdot}5$, which is still very high compared to the typical empirical correlations of order $n^{-1/2} \approx 0{\cdot}015$ that we would expect to observe if the variables were independent.

Our method for variable selection with knockoffs is then applied on the cluster representatives, using the remaining 80% of the observations. The samples used to identify the representatives can also be partially recycled without compromising the rigorous guarantees of false discovery rate control. As shown in Barber & Candès (2016) and Candès et al. (2018), they can be exploited without violating the exchangeability property (2), provided that the corresponding knockoff copies are set equal to the original variables. Alone, these identical knockoffs would not provide any information to distinguish the relevant variables from the nulls.

However, they are useful in improving the accuracy of the feature importance statistics and they can thus increase the power.

Data source	Response	n	p (pre-clustering)	p (post-clustering)
NFBC	HDL (quantitative)	4700	328934	59005
NFBC	LDL (quantitative)	4682	328934	59005
NFBC	TG (quantitative)	4644	328934	59005
NFBC	HT (quantitative)	5302	328934	59005
WTCCC	CD (binary)	4913	377749	71145

Table S3: Summary of the datasets considered in our analysis. The value of n indicates the number of samples for each response, while the last two columns show the corresponding number of variables before and after clustering. Since clustering was performed on the same empirical correlation matrix for all traits in the study of metabolic syndrome, the same number of clusters are found. However, the cluster representatives may be different because they are selected based on the response.

S8. Results of data analysis

S8·1. *Discussion*

The results for the two types of cholesterol are shown in Tables S4 and S5, respectively. In addition to the results in Consortium (2013), we compare our findings to those in Sabatti et al. (2009), an analysis of our same data based on marginal tests with a level of 5×10^{-7}. The latter is different from the canonical $5 \cdot 10^{-8}$ and it was chosen a-posteriori to approximate the threshold obtained with the Benjamini-Hochberg procedure for false discovery rate control at level $\alpha = 0{\cdot}05$. On average, our method makes 8 and $9{\cdot}8$ discoveries for the two types of cholesterol, respectively. These numbers can be compared to the 5 and 6 discoveries reported in Sabatti et al. (2009). For this comparison it should be noted that in Sabatti et al. (2009) several polymorphisms belonging to the same highly correlated region are reported as significant. For the purpose of this comparison, we consider them as one since in our analysis they all belong to the same cluster. In contrast, our procedure rarely selects clusters with overlapping physical positions and we do not further aggregate our findings, because we have already pruned the variables so that variables in different clusters have correlation smaller than $0{\cdot}5$. In Sabatti et al. (2009) an additional association is found within the X chromosome, which we have not analyzed. Among our new findings, some have been confirmed by the meta-analysis in Consortium (2013), while others can be found in the works of different authors. However, we prefer to avoid an extensive search over the entire existing literature to avoid selection bias.

We discover on average $2{\cdot}8$ clusters of polymorphisms associated to triglycerides, as shown in Table S6. This is less than the 4 variants identified in Sabatti et al. (2009), but some of our findings are different and one of the additional ones is confirmed by the meta-analysis.

Height is the last trait from the study of metabolic syndrome that we consider. This is known to be a highly polygenic phenotype, with over 700 known variants. However, the effect of each of these variants is very weak and one should not expect to make many discoveries with a dataset as small as ours. We obtain some validation by comparing our findings to the meta-analyses in Wood et al. (2014); Marouli et al. (2017), as shown in Table S7. Our method discovers 2 relevant clusters, on average. Since this may appear low at first sight, it should be remarked that to the best of our knowledge no other study has found associations for

height using only the data at our disposal. While the longitudinal study in Sovio et al. (2009) has looked for genetic variants associated with height using exclusively this data, none of their reported findings achieves the standard significance threshold.

Our findings on the Crohn's disease data are summarized in Table S8, where we compare them to the meta-analysis in Franke et al. (2010) and the original work of WTCCC (2007). Moreover, we also consider the results of Candès et al. (2018), whose work is the most similar to ours because it uses the same data, pre-processing and clustering method, as well as the overall knockoff methodology. The important distinction is that they construct their knockoff variables differently. Instead of fitting a hidden Markov model to the genetic sequences, they assume a multivariate normal distribution. Their nominal false discovery rate target $\alpha = 0{\cdot}1$ is the same as ours, and WTCCC (2007) also aims at controlling the Bayesian false discovery rate at approximately the same level. Our method makes $22{\cdot}8$ discoveries on average, versus 18 in Candès et al. (2018) and the 9 of WTCCC (2007). In addition to an apparently higher power in this case, our procedure can in general be expected to enjoy a more principled and safer guarantee. Nowhere have we made the unrealistic assumptions of WTCCC (2007) on the conditional model for the response nor those of Candès et al. (2018) on the model for the covariates. Several of the additional findings that we make have been confirmed in Franke et al. (2010), as shown in Table S8.

S8·2. *Tables*

We report below the findings of our data analysis performed on the five phenotypes considered in this paper. An asterisk indicates the presence of a confirmed association within $0{\cdot}5$ mega base pairs of our discovered cluster. We also compute marginal p-values with the standard univariate analysis for all selected polymorphisms and show the smallest one in each cluster. It must be remarked that our p-values are not identical to those in the original studies, since we have made slightly different methodological choices in the pre-processing and pruning phases, as detailed in § 7·1. It interesting to look at these p-values because they highlight that many of the marginal signals are weak and could not have been detected by a traditional procedure.

S8·3. *HDL cholesterol*

Selection frequency	SNP (cluster size)	Chr.	Position range (Mb)	Confirmed in Consortium (2013)	Found in Sabatti et al. (2009)	Marginal p-value
100%	rs1532085 (4)	15	58·68–58·7	rs1532085	rs1532085	$1{\cdot}33 \cdot 10^{-12}$
100%	rs7499892 (1)	16	57·01–57·01	rs3764261	rs3764261	$9{\cdot}55 \cdot 10^{-17}$
100%	rs1800961 (1)	20	43·04–43·04	rs1800961		$2{\cdot}84 \cdot 10^{-8}$
99%	rs1532624 (2)	16	56·99–57·01	rs3764261	rs3764261	$3{\cdot}08 \cdot 10^{-34}$
95%	rs255049 (142)	16	66·41–69·41	rs16942887	rs255049	$1{\cdot}76 \cdot 10^{-08}$
57%	rs10096633 (19)	8	19·73–19·94			$5{\cdot}33 \cdot 10^{-06}$
55%	rs9898058 (1)	17	47·82–47·82			$1{\cdot}43 \cdot 10^{-06}$
51%	rs17075255 (59)	5	164·28–164·92			$1{\cdot}38 \cdot 10^{-05}$
43%	rs3761373 (1)	21	42·87–42·87			$5{\cdot}96 \cdot 10^{-06}$
28%	rs2575875 (10)	9	107·63–107·68	rs3905000		$1{\cdot}04 \cdot 10^{-06}$
23%	rs12139970 (11)	1	230·35–230·42	rs4846914		$1{\cdot}21 \cdot 10^{-05}$
12%	rs173738 (3)	5	16·71–16·73			$4{\cdot}77 \cdot 10^{-06}$

Table S4: Clusters of polymorphisms found to be associated with HDL cholesterol over 100 repetitions of our procedure. Positions follow the convention of the Human Genome Build 37, as in the original data. The marginal p-values are obtained from standard univariate linear regression.

S8·4. *LDL cholesterol*

Selection frequency	SNP (cluster size)	Chr.	Position range (Mb)	Confirmed in Consortium (2013)	Found in Sabatti et al. (2009)	Marginal p-value
99%	rs4844614 (34)	1	207·3–207·88		rs4844614	$2{\cdot}00 \cdot 10^{-9}$
97%	rs646776 (5)	1	109·8–109·82	rs629301	rs646776	$2{\cdot}49 \cdot 10^{-9}$
97%	rs2228671 (2)	19	11·2–11·21	rs6511720	rs11668477	$2{\cdot}28 \cdot 10^{-9}$
94%	rs157580 (4)	19	45·4–45·41	rs4420638*	rs157580	$3{\cdot}62 \cdot 10^{-8}$
92%	rs557435 (21)	1	55·52–55·72	rs2479409		$1{\cdot}17 \cdot 10^{-7}$
80%	rs10198175 (1)	2	21·13–21·13	rs1367117*	rs693*	$5{\cdot}05 \cdot 10^{-7}$
76%	rs10953541 (58)	7	106·48–107·3			$3{\cdot}75 \cdot 10^{-6}$
62%	rs6575501 (1)	14	95·64–95·64			$2{\cdot}32 \cdot 10^{-6}$
41%	rs1713222 (45)	2	21·11–21·53	rs1367117	rs693	$4{\cdot}99 \cdot 10^{-11}$
40%	rs2802955 (1)	1	235·02–235·02	rs514230*		$2{\cdot}27 \cdot 10^{-1}$
37%	rs17129799 (23)	11	96·85–97			$4{\cdot}84 \cdot 10^{-6}$
36%	rs174450 (16)	11	61·55–61·68	rs174546	rs1535	$9{\cdot}96 \cdot 10^{-7}$
26%	rs905502 (1)	8	3·13–3·13			$1{\cdot}30 \cdot 10^{-4}$
25%	rs9696070 (6)	9	89·21–89·24			$1{\cdot}26 \cdot 10^{-5}$
23%	rs166152 (19)	16	29·04–29·33			$4{\cdot}29 \cdot 10^{-5}$
19%	rs12427378 (43)	12	50·43–51·31			$3{\cdot}69 \cdot 10^{-6}$

Table S5: Clusters of polymorphisms found to be associated with LDL cholesterol. Other details as in caption of Table S4.

S8·5. *Triglycerides*

Selection frequency	SNP (cluster size)	Chr.	Position range (Mb)	Confirmed in Consortium (2013)	Found in Sabatti et al. (2009)	Marginal p-value
94%	rs10096633 (19)	8	19·73–19·94	rs12678919	rs10096633	$7{\cdot}47 \cdot 10^{-8}$
91%	rs676210 (45)	2	21·11–21·53		rs673548	$2{\cdot}00 \cdot 10^{-7}$
62%	rs2304130 (37)	19	19·28–19·87	rs10401969		$3{\cdot}91 \cdot 10^{-6}$
25%	rs2907632 (13)	17	52·86–52·95			$5{\cdot}69 \cdot 10^{-6}$

Table S6: Clusters of polymorphisms found to be associated with triglycerides. Other details as in caption of Table S4.

S8·6. *Height*

Selection frequency	SNP (cluster size)	Chr.	Position range (Mb)	Confirmed in Wood et al. (2014)	Confirmed in Marouli et al. (2017)	Marginal p-value
68%	rs2814982 (120)	6	34·17–35·45	rs12214804	rs2814982	$1{\cdot}33 \cdot 10^{-7}$
46%	rs2882676 (5)	15	89·39–89·4		rs2882676	$2{\cdot}73 \cdot 10^{-6}$

31%	rs6763931 (14)	3	141·04–141·34	rs724016	rs724016*	$4{\cdot}00 \cdot 10^{-6}$
12%	rs10769671 (17)	11	6·19–6·28			$6{\cdot}37 \cdot 10^{-6}$

Table S7: Clusters of polymorphisms found to be associated with height. Other details as in caption of Table S4.

S8·7. *Crohn's disease*

Selection frequency	SNP (cluster size)	Chr.	Position range (Mb)	Confirmed in Franke et al. (2010)	Found in WTCCC (2007)	Found in Candès et al. (2018)	Marginal p-value
100%	rs11209026 (2)	1	67·31–67·42	rs11209026	rs11805303	100%	$2{\cdot}57 \cdot 10^{-21}$
99%	rs6431654 (20)	2	233·94–234·11	rs3792109	rs10210302	100%	$1{\cdot}44 \cdot 10^{-14}$
98%	rs6688532 (33)	1	169·4–169·65		rs12037606	90%	$3{\cdot}48 \cdot 10^{-8}$
97%	rs17234657 (1)	5	40·44–40·44	rs11742570	rs17234657	90%	$8{\cdot}06 \cdot 10^{-13}$
95%	rs11805303 (16)	1	67·31–67·46	rs11209026	rs11805303	100%	$5{\cdot}22 \cdot 10^{-14}$
91%	rs7095491 (18)	10	101·26–101·32	rs4409764	rs10883365	100%	$2{\cdot}81 \cdot 10^{-7}$
91%	rs3135503 (16)	16	49·28–49·36	rs2076756	rs17221417	90%	$9{\cdot}55 \cdot 10^{-11}$
81%	rs7768538 (1145)	6	25·19–32·91	rs1799964	rs9469220	60%	$5{\cdot}83 \cdot 10^{-9}$
80%	rs6601764 (1)	10	3·85–3·85		rs6601764	100%	$1{\cdot}83 \cdot 10^{-8}$
75%	rs7655059 (5)	4	89·5–89·53			40%	$2{\cdot}14 \cdot 10^{-7}$
73%	rs6500315 (4)	16	49·03–49·07	rs2076756	rs17221417	60%	$5{\cdot}73 \cdot 10^{-7}$
72%	rs2738758 (5)	20	61·71–61·82	rs4809330		60%	$2{\cdot}64 \cdot 10^{-6}$
70%	rs7726744 (46)	5	40·35–40·71	rs11742570	rs17234657	50%	$7{\cdot}24 \cdot 10^{-13}$
68%	rs11627513 (7)	14	96·61–96·63			80%	$6{\cdot}70 \cdot 10^{-6}$
66%	rs4246045 (46)	5	150·07–150·41	rs7714584	rs1000113	50%	$2{\cdot}00 \cdot 10^{-8}$
62%	rs9783122 (234)	10	106·43–107·61			80%	$1{\cdot}69 \cdot 10^{-4}$
61%	rs6825958 (3)	4	55·73–55·77			30%	$3{\cdot}54 \cdot 10^{-5}$
56%	rs4692386 (1)	4	25·81–25·81			40%	$1{\cdot}31 \cdot 10^{-6}$
56%	rs4263839 (23)	9	114·58–114·78			30%	$3{\cdot}16 \cdot 10^{-5}$
54%	rs2390248 (13)	7	19·8–19·89			50%	$4{\cdot}53 \cdot 10^{-7}$
51%	rs10916631 (14)	1	220·87–221·08			40%	$5{\cdot}41 \cdot 10^{-5}$
49%	rs4437159 (4)	3	84·8–84·81			60%	$5{\cdot}42 \cdot 10^{-5}$
48%	rs9469615 (2)	6	33·91–33·92			30%	$1{\cdot}13 \cdot 10^{-5}$
45%	rs10761659 (53)	10	64·06–64·41	rs10761659	rs10761659	10%	$2{\cdot}55 \cdot 10^{-6}$
42%	rs2836753 (5)	21	39·21–39·23			30%	$1{\cdot}43 \cdot 10^{-6}$
39%	rs6743984 (23)	2	230·91–231·05	rs7423615		10%	$3{\cdot}79 \cdot 10^{-6}$
38%	rs2279980 (20)	5	57·95–58·07			10%	$1{\cdot}08 \cdot 10^{-6}$
35%	rs7186163 (6)	16	49·2–49·25	rs2076756	rs17221417	50%	$7{\cdot}29 \cdot 10^{-8}$
32%	rs16857006 (1)	2	11·1–11·1				$2{\cdot}30 \cdot 10^{-3}$
30%	rs7807268 (5)	7	147·65–147·7			10%	$2{\cdot}57 \cdot 10^{-5}$
27%	rs4807569 (2)	19	1·07–1·08	rs740495			$2{\cdot}06 \cdot 10^{-5}$
24%	rs3779585 (2)	7	90·36–90·38				$7{\cdot}40 \cdot 10^{-6}$
23%	rs12529198 (31)	6	5·01–5·1				$1{\cdot}08 \cdot 10^{-6}$
22%	rs7497036 (19)	15	72·49–72·73				$2{\cdot}04 \cdot 10^{-4}$
20%	rs4959830 (11)	6	3·36–3·41	rs17309827		10%	$9{\cdot}47 \cdot 10^{-7}$
15%	rs13282050 (8)	8	69·3–69·31				$3{\cdot}64 \cdot 10^{-5}$
15%	rs1451890 (26)	15	30·92–31·01				$1{\cdot}23 \cdot 10^{-5}$
14%	rs2814036 (5)	1	163·94–164·07				$9{\cdot}31 \cdot 10^{-7}$
14%	rs7759649 (2)	6	21·57–21·58	rs6908425*		40%	$1{\cdot}01 \cdot 10^{-4}$
14%	rs4870943 (10)	8	126·59–126·62	rs4871611			$1{\cdot}46 \cdot 10^{-6}$
11%	rs10923347 (1)	1	117·83–117·83				$9{\cdot}54 \cdot 10^{-4}$
10%	rs4438299 (30)	16	60·01–60·32				$7{\cdot}07 \cdot 10^{-5}$

Table S8: Clusters of polymorphisms found to be associated with Crohn's disease over 100 repetitions of our procedure. Positions follow the convention of the Human Genome Build 35, as in the original data. The marginal p-values are obtained from the Cochran-Armitage test for trend (Armitage, 1955).

S9. GLOSSARY OF GENETIC TERMS

For those readers who may not be familiar with some of the genetic terms used in this paper, we provide a short glossary. The definitions below refer to the use of the terms within this paper and may not correspond exactly with the most general definition of the terms outside our context.

Allele: one of two nucleotides at a precise position on the DNA, inherited from one parent.

Chromosome: in general, a sequence of DNA forming a single molecule. The complete DNA of an individual is divided into 23 pairs of chromosomes. Here, we refer to a chromosome as one such pair, thus representing a sequence of genotypes inherited by both parents.

Gene: a sequence of loci that code a specific function, i.e. control one or more phenotypes.

Genome-wide association study: an observational study of a set of single-nucleotide polymorphism in different individuals aimed at identifying variant that are associated with a specific phenotype.

Genotype: a pair of alleles at a precise position on the DNA, each belonging to a different haplotype and inherited from one parent.

Haplotype: a sequence of alleles on different loci, inherited from a single parent.

Hardy-Weinberg equilibrium: a basic principle of population genetics, stating that the amount of genetic variation in a population will remain constant across generations in the absence of disturbing factors such as mutation, selection or mate choice based on desired phenotypes. In genome-wide association studies, it is customary to verify that the genotype frequences are consistent with this principle in order to detect possible genotyping errors, batch effects or population stratification.

Locus: a fixed position on a chromosome, indicating a single-nucleotide polymorphism.

Marker: in general, a short DNA sequence at a known locus, used to study the relationship between an inherited disease and its genetic cause. Here, it refers to a single-nucleotide polymorphism and it is used interchangeably.

Phase: the original combination of alleles that an individual inherited from its parents. Typically, in genome-wide association studies the genotypes are observed as unordered pairs of alleles and the individual haplotypes not directly accessible.

Phenotype: an observable characteristics of an individual, resulting from the expression of the genetic code as well as the influence of environmental factors.

Nucleotide: an organic molecule serving as the basic structural unit of the DNA, either adenine, thymine, cytosine, or guanine.

Recombination rate: the frequency of recombination events, in which a portion of the DNA of the offspring differs from those of both parents as a result of an error during cell division.

Single-nucleotide polymorphism: a precise position on the DNA where a single nucleotide can differ between people. Two alleles are found at each such position. It is the marker of choice in genome-wide association studies. It is sometimes referred to in short as a polymorphism or abbreviated as SNP.

Variant: a genetic feature that varies between different individuals, here understood to be a single-nucleotide polymorphism.

REFERENCES

ARMITAGE, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics* **11**, 375–386.

BARBER, R. F. & CANDÈS, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.* **43**, 2055–2085.

BARBER, R. F. & CANDÈS, E. J. (2016). A knockoff filter for high-dimensional selective inference. *arXiv:1602.03574* .

CANDÈS, E. J., FAN, Y., JANSON, L. & LV, J. (2018). Panning for gold: "model-X" knockoffs for high dimensional controlled variable selection. *J. R. Statistic. Soc. B* **80**, 551–577.

CAWLEY, S. L. & PACHTER, L. (2003). HMM sampling and applications to gene finding and alternative splicing. *Bioinformatics* **19**, ii36–ii41.

CONSORTIUM, G. L. G. (2013). Discovery and refinement of loci associated with lipid levels. *Nature Genet.* **45**, 1274–1283.

FRANKE, A., MCGOVERN, D. P. B., BARRETT, J. C., WANG, K., RADFORD-SMITH, G. L., AHMAD, T. & ET AL. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genet.* **42**, 1118–1125.

JUANG, B. H. & RABINER, L. R. (1991). Hidden Markov models for speech recognition. *Technometrics* **33**, 251–272.

MAROULI, E., GRAFF, M., MEDINA-GOMEZ, C., LO, K. S., WOOD, A. R., KJAER, T. R., FINE, R. S., LU, Y., SCHURMANN, C., HIGHLAND, H. M. et al. (2017). Rare and low-frequency coding variants alter human adult height. *Nature* **542**, 186–190.

PACIFICO, M. P., GENOVESE, C., VERDINELLI, I. & WASSERMAN, L. (2004). False discovery control for random fields. *J. Am. Statist. Assoc.* **99**, 1002–1014.

PRICE, A. L., PATTERSON, N. J., PLENGE, R. M., WEINBLATT, M. E., SHADICK, N. A. & REICH, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genet.* **38**, 904–909.

RABINER, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77**, 257–286.

SABATTI, C., HARTIKAINEN, A.-L., POUTA, A., RIPATTI, S., BRODSKY, J., JONES, C. G., ZAITLEN, N. A., VARILO, T., KAAKINEN, M., SOVIO, U. et al. (2009). Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature Genet.* **41**, 35–46.

SIEGMUND, D. O., ZHANG, N. R. & YAKIR, B. (2011). False discovery rate for scanning statistics. *Biometrika* **98**, 979.

SOVIO, U., BENNETT, A. J., MILLWOOD, I. Y., MOLITOR, J., O'REILLY, P. F., TIMPSON, N. J. & ET AL. (2009). Genetic determinants of height growth assessed longitudinally from infancy to adulthood in the northern finland birth cohort 1966. *PLOS Genet.* **5**, e1000409.

WOOD, A. R., ESKO, T., YANG, J., VEDANTAM, S., PERS, T. H., GUSTAFSSON, S., CHU, A. Y., ESTRADA, K., LUAN, J., KUTALIK, Z. et al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genet.* **46**, 1173–1186.

WTCCC (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678.

ZHU, J., LIU, J. S. & LAWRENCE, C. E. (1998). Bayesian adaptive sequence alignment algorithms. *Bioinformatics* **14**, 25.