Sequence analysis

**KIMI: Knockoff Inference for Motif Identification from molecular sequences with controlled false discovery rate**

Xin Bai\(^1\), Jie Ren\(^1\), Yingying Fan\(^2,\ast\) and Fengzhu Sun\(^1,\ast\)

\(^1\)Quantitative and Computational Biology Program, Department of Biological Sciences, University of Southern California, Los Angeles, CA, 90089, USA.  
\(^2\)Data Sciences and Operations Department, Marshall School of Business, University of Southern California, Los Angeles, CA, 90089, USA.

\(\ast\) To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

**Abstract**

**Motivation:** The rapid development of sequencing technologies has enabled us to generate a large number of metagenomic reads from genetic materials in microbial communities, making it possible to gain deep insights into understanding the differences between the genetic materials of different groups of microorganisms, such as bacteria, viruses, plasmids, etc. Computational methods based on \(k\)-mer frequencies have been shown to be highly effective for classifying metagenomic sequencing reads into different groups. However, such methods usually use all the \(k\)-mers as features for prediction without selecting relevant \(k\)-mers for the different groups of sequences, i.e, unique nucleotide patterns containing biological significance.

**Results:** To select \(k\)-mers for distinguishing different groups of sequences with guaranteed false discovery rate (FDR) control, we develop KIMI, a general framework based on model-X Knockoffs regarded as the state-of-the-art statistical method for false discovery rate (FDR) control, for sequence motif discovery with arbitrary target FDR level, such that reproducibility can be theoretically guaranteed. KIMI is shown through simulation studies to be effective in simultaneously controlling FDR and yielding high power, outperforming the broadly used Benjamini-Hochberg (B-H) procedure and the \(q\)-value method for FDR control. To illustrate the usefulness of KIMI in analyzing real datasets, we take the viral motif discovery problem as an example and implement KIMI on a real dataset consisting of viral and bacterial contigs. We show that the accuracy of predicting viral and bacterial contigs can be increased by training the prediction model only on relevant \(k\)-mers selected by KIMI.

**Availability:** Our implementation of KIMI is available at https://github.com/xinbaiusc/KIMI.

**Contact:** fanyingy@marshall.usc.edu or fsun@usc.edu

**Supplementary information:** Supplementary Materials are available at Bioinformatics online.

---

1 Introduction

With the ubiquitous data available in the era of big data, the identification of key variables that are relevant to an effect has become tremendously important but challenging because of the potentially complicated dependance structure such as the nonlinearity, high dimensionality and collinearity. Yet, despite the large body of literature on variable selection, reproducibility of the results still remains challenging, especially when dealing with biological data that often exhibit significant variation (Sinha et al., 2017; Ricós et al., 2007).

Among all attempts to address the reproducibility issue, measuring the false discovery rate (FDR) has been widely used. Existing methods on FDR control typically depend on calculating \(p\)-values measuring the significance of variables, such as the very widely...
used Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995). One basic assumption of the BH procedure is that valid
\( p \)-values satisfying additional conditions such as independence or positive dependencies can be calculated for testing the variable
significance. However, this can be difficult to guarantee especially in nonlinear or high dimensional cases. Recently, Candes et al.
(2018) proposed a new framework of model-X knockoffs to control the FDR bypassing the use of \( p \)-values. The salient idea of model-X
knockoffs is to create the so-called knockoff variables that mimic the dependence of original variables but are irrelevant to the response
conditional on the original variables. Thus, knockoff variables can act as controls for assessing importance of original variables. It is
proved therein that model-X knockoffs can control FDR at the tar-
get level under arbitrary dimensionality and arbitrary dependence
structure between covariates and the response.

Due to the rapid development of sequencing technologies, innumerable metagenomic reads have been generated, while the
understanding of the underlying principles governing the assembly of microbial communities remains far behind. Recent studies fou-
sing on the microbial community have provided us deep insights into the relationship between microbial communities and human dis-
ases (Dethlefsen et al., 2007), the environment (Hamman et al., 2007), and the biosphere (Sogin et al., 2006), etc. Different groups of
microorganisms, such as bacteria, viruses, plasmids, etc, play distinct roles and interact with each other in the microbial commu-
nity. Due to the nature of shotgun sequencing, the sources of the short reads are lost. One basic yet essential problem is to classify
sequences into different categories in mixed metagenomic data-
sets. Several computational tools have been developed to solve this
problem (Ren et al., 2020; Roux et al., 2015; Ren et al., 2017; Fang et al., 2019). Among all these efforts, VirFinder (Ren et al.,
2017), a novel \( k \)-mer based classification tool, uses frequencies of all \( k \)-mers (or \( k \)-tuples, \( k \)-grams) as features to build the model for
classifying viruses and bacteria. Although high accuracy can be achieved by VirFinder for the prediction of viral contigs, it remains
important to know which \( k \)-mers are indeed uniquely enriched or
depleted in viruses. While logistic regression with lasso penalty can produce some support recovery, it is likely to miss some impor-
tant ones or to have inflated FDR. Therefore, the understanding
towards the scientific problem of the accurate identification of viral
specific motifs, i.e., \( k \)-mers that have different abundances in viru-
ses and bacteria, is lacking. The identification of specific \( k \)-mers
containing particular biological significance to a group of species, also
named motifs considered as patterns of residues or regions
within nucleotide sequences, is a key part of understanding function
and regulation within biological systems (Lones and Tyrrell, 2005).

Taking viruses as an example, understanding viral motifs could not
only help us identify viral sequences more accurately, but also pro-
mote the understanding of virus-host interactions, in which shuffled
motifs help to evade clustered regularly interspaced short palindro-
mic repeats (CRISPR) spacers (Andersson and Banfield, 2008).

Despite the fact that many existing studies are available for motif
finding (Galas et al., 1985; Mengeritsky and Smith, 1987; Lawre-
nce and Reilly, 1990; Lawrence et al., 1993; Marsan and Sagot, 2000), none of them could simultaneously control the FDR.

In this paper, we start from comparing the occurrence frequen-
cies of \( k \)-mers in two types of sequences given the sequence
labels. Our target is to identify important motifs (\( k \)-mers) that are
relevant to the binary sequence types with controlled FDR. One
challenge is that the frequencies of all \( k \)-mers, as compositional
data, sum up to one, resulting in the covariates being perfectly
collinear. A common practice to overcome this problem is to drop
one \( k \)-mer. However, this does not completely solve our problem
because the remaining \( k \)-mer frequencies can still have very high
collinearity. Another natural way to deal with compositional covaria-
tes is to use the log-ratio transformed covariates, which is equivalent
to taking log of the covariates with the constraint that the coeffici-
ents of the covariates sum up to 0, as proposed in (Lin et al., 2014).

We adopt the former one for KIMI throughout the paper and com-
pare the performance of KIMI with the results based on the log-ratio
transformation in the real data analysis. The second challenge is
that the joint distribution of covariates is assumed to be known for
the construction of knockoff variables, which specifically means the
joint distribution of all \( k \)-mer frequencies need to be known. Thus,
besides the afore-mentioned issue of high collinearity, we also face
the problem of unknown covariate distribution. These challenges
indicate that it is difficult to naively adapt the model-X knockoffs
framework to our motif identification problem.

We present KIMI, a general framework based on knockoff
inference, for motif identification from binary types of molecular
sequences with FDR control. First, we briefly review the framework
of model-X knockoffs. The key step for FDR control and for overco-
ming the collinearity issue is to generate valid knockoff variables
adapting to our problem. To tackle this, we directly generate the
knockoff variables for \( k \)-mer frequencies of the original sequences
by assuming multivariate normal distribution, which is shown to be
able to control FDR and obtain high power simultaneously in simula-
tion studies. The widely used BH procedure and the \( \gamma \)-value method
for FDR control, on the contrary, yielded almost no power. Finally,
we consider an application example of viral motif identification,
in which we test KIMI on real viral and bacterial contigs from VirFinder
(Ren et al., 2017). Our results show that higher prediction accuracy
can be obtained by training the prediction model only on relevant
\( k \)-mers selected by KIMI, although predicting the contig labels is
not our primary goal. This application example demonstrates the
promising use of KIMI for analyzing real datasets.

2 Methods

2.1 Counting \( k \)-mer frequencies from metagenomic contigs

Metagenomic reads are generally short consisting of several hun-
dred of basepairs, making it challenging for statistical analysis. The
first step in most metagenomic studies is to assemble the reads into
contigs, consecutive regions of genomes with overlapping reads. To
facilitate statistical analysis, we consider contigs with length above
a certain threshold. Suppose there are \( n \) contigs, each labeled as
\( y_{i} = 1 \) or \( y_{i} = 0, 0.1 \leq i \leq n \) corresponding to two different cate-
gories, for example, viral or bacterial contigs. Our objective is to
develop a method that can classify whether an observed contig
comes from a virus or bacterium and simultaneously identify the
key factors driving the successful classification. We will consider
similar problems for multiple type classification in future studies.

We represent the \( i \)-th contig as \( Z_{i} = Z_{i,1}Z_{i,2} \cdots Z_{i,L} \), where
\( Z_{i,j} \in A = \{A,C,G,T\} \) denotes the \( j \)-th nucleotide in the \( i \)-
th contig and \( i \) stands for the contig length. For any word \( w =
\underbrace{w_{1}w_{2} \cdots w_{l}}_{l \text{ mers of length } l} \), we count its number of occurrences in con-
tig \( Z_{i} \), and denote it as \( N_{i}(w) \). Then the \( k \)-mer counts are further
normalized into frequencies denoted as \( f_{j}(w) \) by dividing the total
counts of all \( k \)-mers. Some previous studies used logistic regres-
sion on \( k \)-mers (Akhter et al., 2012; Ren et al., 2017) to distinguish
viral contigs from bacterial ones, showing strong associations bet-
ween the composition abundance of \( k \)-mers and contig labels. In this
study, we focus on identifying \( k \)-mers that are particularly relevant
to binary types of sequences, i.e., motifs containing essential biological meanings for characterizing particular types of sequences. Hereafter, to simplify the presentation, we use $S_0$ to denote the set of such $k$-mers and refer them as relevant $k$-mers.

For any contig $Z$, the $k$-mer frequencies always add up to one, i.e., $\sum_{w} f_i(w) = 1$. Thus, naively including all $k$-mer frequencies will lead to the problem of perfect collinearity in many computational methods. Thus, to alleviate the collinearity problem, we propose to drop one $k$-mer with the least likelihood of being relevant to viral contig identification according to some criterion. The details for selecting the dropped $k$-mer will be discussed later in Sections 2.4 and 2.5. Thus, the total number of interested $k$-mers will be $p = 4^k - 1$.

Note that even after dropping one $k$-mer, the collinearity among $k$-mer frequencies can still be very high, which is indeed the main challenge in our study.

Hereafter to simplify notation, we use $x_i = (X_{i1}, \cdots, X_{i\rho})' \in \mathbb{R}^\rho$, $i = 1, \cdots, \eta$ to denote the vectors of transformed $k$-mer frequencies (the transformation will be specified later), and $y_i$'s to denote the corresponding contig labels. The problem of identifying the sets of useful $k$-mers $S_{0k}$ from the full set of candidates is termed as variable selection in statistics literature. We aim at recovering $S_{0k}$ from observed $(x_i, y_i) = (1, \cdots, \eta)$ with controlled false discovery rate.

### 2.2 The knockoffs framework for variable selection and FDR control

To evaluate variable selection methods, we first introduce some performance measures. Let $S$ be the set of $k$-mers selected by some statistical methods. The associated false discovery rate (FDR) can be defined as:

$$FDR = \mathbb{E}[\text{FDP}], \quad \text{with } \text{FDP} = \left| \frac{\left| S \setminus S_0 \right|}{|S|} \right|,$$

where FDP means false discovery proportion, $\cdot$ denotes the cardinality of a set, and $S_0$ is the complement of $S$, that is, the set of non-relevant $k$-mers. Correspondingly, the power, which is the expected fraction of selected relevant $k$-mers, is defined as:

$$\text{Power} = \mathbb{E}[\text{TP}], \quad \text{with } \text{TP} = \left| \frac{\left| S_0 \cap S \right|}{|S_0|} \right|,$$

where TP means true discovery proportion. We aim to control FDR and achieve high power simultaneously for the motif discovery problem.

In Candes et al. (2018), a general model-X knockoffs framework was proposed to control FDR, which can be regarded as a wrapper and can be combined with any underlying variable selection methods that assign variable importance measure to achieve FDR control. It has been proved in Candes et al. (2018) that the model-X knockoffs framework controls FDR at the desired level with finite sample size for any dependence structure between the response and predictors. It also automatically adapts to the collinearity level in predictors. Therefore, we start from the idea of model-X knockoffs and adapt it to select relevant $k$-mers. We next give a brief review on the Model-X knockoffs framework.

The salient idea of model-X knockoffs can be intuitively understood as creating a set of “fake” variables, the so-called knockoff variables, that mirror the dependence structure of original variables, but are irrelevant to the dependent variable $Y$ conditional on the original variables. These knockoff variables are then used as controls for assessing importance of the original variables. The formal definition of Model-X knockoff variables (Candes et al., 2018) is given below:

Definition 1. (Candes et al., 2018) Model-X knockoffs for the family of random variables $x = (X_1, \cdots, X_\rho)'$ is a new family of random variables $\tilde{x} = (\tilde{X}_1, \cdots, \tilde{X}_\rho)'$, constructed such that:

- For any subset $S \subset \{1, \cdots, p\}$
  $$\tilde{X}_j \overset{d}{=} (X_j, \tilde{X}_j)_\text{swap(S)},$$

- Conditional on the original variables $x$, the knockoff variables $\tilde{x}$ are independent of the response $Y$.

Hereafter to simplify notation, we use $x_i = (X_{i1}, \cdots, X_{i\rho})' \in \mathbb{R}^\rho$, $i = 1, \cdots, \eta$ and $\tilde{y}_i$ means that the two vectors have identical distribution.

Candes et al. (2018) suggested a general algorithm for generating knockoff variables when the joint distribution of $x$ is known. We will discuss the details of implementation in our setting with unknown covariate distribution in Section 2.3 for ease of presentation.

Now we assume such valid Model-X knockoffs have been generated. Let $\tilde{x}_i$ be the knockoff variable of $x_i$, and $x_i' = (x_i', \tilde{x}_i)' \in \mathbb{R}^{2\rho}$ the augmented feature vector. With $(\tilde{x}_i', \tilde{y}_i')$, $i = 1, \cdots, \eta$, we fit the following regularized logistic regression with the L1 penalty:

$$\min_{b \in \mathbb{R}^{2\rho}} \left\{ -\frac{1}{\eta} \sum_{i=1}^{\eta} \left[ y_i \log (1 + \exp(b' \tilde{x}_i')) - \log (1 + \exp(b' \tilde{y}_i' + b' \tilde{x}_i')) \right] + \lambda \|b\|_1 \right\},$$

where $\lambda > 0$ is the regularization parameter. In implementation, we choose $\lambda$ by the $K$-fold cross validation method based on prediction error.

With the obtained solution $b = (b_1, \cdots, b_\rho)'$ to problem (3), we then calculate the difference between the magnitudes of coefficients of the original variables and their corresponding knockoff copies, that is,

$$V_j = |b_j| - |b_{j+\rho}|, \quad j = 1, \cdots, \rho.$$
transition matrices. One can then calculate $k$-mer frequencies from the original contigs and knockoff contigs. However, this approach does not produce valid knockoffs for $k$-mer frequencies because the $k$-mer frequencies from knockoff contigs are not exchangeable in distribution with the ones from the original contigs. Due to page limitations, we discuss the knockoff Markov contig method and provide detailed explanation on why it fails in our study in the Supplementary Materials. Instead of modeling knockoff contigs, we directly generate knockoff $k$-mer frequencies by assuming multivariate normal distribution of $k$-mer frequencies.

2.3 Generating knockoff $k$-mer frequencies

An applicable approach is to directly generate knockoffs for $k$-mer frequencies, considering that our goal is to select relevant $k$-mer frequencies. Let $F = (f_1, \ldots, f_p)$, each $f_i$ is a vector of $k$-mer frequencies, in which $n$ is the number of samples, $k$ is the total number of $k$-mers of interest after dropping one $k$-mer. Then rows of $F$, denoted by $f_1, \ldots, f_p$, can be regarded as an estimate of multivariate normal probabilities, where each entry of the multinomial distribution represents one $k$-mer. Classical central limit theory motivates us to model the rows of $F$ as independent normal random vectors.

We further standardize each column of $F$ to have mean 0 and standard deviation 1, and denote by $X = (x_1, \ldots, x_p) \in \mathbb{R}^{n \times p}$ the $k$-mer frequency matrix after standardization. Thus, the rows of $X$ all have mean 0 and covariance matrix with diagonals equal 1. In addition, we assume that the distribution of rows of $X$ is still close to multivariate normal. Assume indeed $x_i \sim \mathcal{N}(0, \Sigma_{x_i})$ where $\text{diag}(\Sigma_{x_i}) = I_p$. Then a valid construction of knockoff variables is

$$x_i \mid x_i \sim N(\hat{x}_i - \text{diag}(x_i) \beta, 2 \text{diag}(x_i) - \text{diag}(x_i) \text{diag}(x_i)),$$

where $\Omega = \Sigma_{x_i}^{-1}$ is the $p \times p$ precision matrix (the inverse covariance matrix), and $\beta$ is a vector of $p$ nonnegative numbers satisfying $2\text{diag}(\beta) - \text{diag}(x_i)\text{diag}(\beta)$ being positive definite. Here $\text{diag}(x_i)$ is a diagonal matrix with items of $x_i$ on the diagonal. The vector $\beta$ measures the dissimilarity between the original variables and the knockoff variables. We denote the knockoff $k$-mer frequency matrix by $X' = (x'_1, \ldots, x'_p)$. Notice that due to the high dimensionality of the data, the normality assumption may be miss-specified and the constructed knockoff variables are not ideal. However, we evaluate KIMI by its variable selection and prediction performance, and show that it achieves controlled FDR and high power simultaneously in simulation studies.

2.4 Simulation studies

2.4.1 Simulation procedures for FDR control using KIMI

We start from simulating contigs using Markov models given that Markov chains have been widely used in molecular sequence analyses (Almogi, 1983; Arnold et al., 1988; Avery, 1987; Avery and Henderson, 1999; Blaisdell, 1986, 1985; Reinert et al., 2000; Waterman, 1995). Note that the Markov model assumption is only needed to facilitate simulation studies and is not essential since no assumptions on the data are necessary for KIMI. We first generate $n$ Markov contigs with order $r$ and length $L$. The contigs are simulated from the same model and their labels will be determined later by prespecified relevant $k$-mers. The transition probability matrix of Markov contigs $T$ with dimension $4^r \times 4^r$ is randomly simulated as follows. For each row of $T$, we sample $r$ random numbers following a uniform distribution in $(0,1)$ and then normalize by the sum of these $r$ numbers assuring the row sum equals to 1. These normalized numbers are treated as the row entries of $T$. We repeat the random sampling until all rows of $T$ are generated. After that, we simulate Markov contigs by starting from the initial distribution of equal probabilities $(1/4^r, \ldots, 1/4^r)$. Then we continue the simulation according to the corresponding transition probability matrix until the contig length reaches $L$.

We count the occurrence of each $k$-mer in every original contig and normalize the counts into frequency. We further standardize the $k$-mer frequency matrix $F$ as described at the end of Section 2.3, and denote by $X$ the final data matrix.

Next, we sample the binary contig labels according to the following procedures. First, we randomly pick $k$-mers from all $4^r$ $k$-mers to be relevant. Second, we sample $y$ from a Bernoulli distribution with parameter based on the following logistic model

$$P(y_{ij} = 1) = \frac{\exp(x'_i \beta_j)}{1 + \exp(x'_i \beta_j)},$$

where $\beta = (\beta_1, \ldots, \beta_{4^r})$ is a coefficient vector corresponding to the effect of each $k$-mer on the contig labels. For those irrelevant $k$-mers, we assign 0 as the coefficient. Those relevant $k$-mers will be equally likely assigned $a$ or $-a$ as the coefficient, where $a$ is a positive number indicating the amplitude of contribution to the probability of having label 1. Finally, we use the Wilcoxon-Mann-Whitney (WMW) test (Mann and Whitney, 1947) to compare each column of $X$ for the two types of contigs and drop the $k$-mer having the largest $p$-value, i.e., least likely being a relevant $k$-mer. A knockoff matrix $X$ is then constructed using equation (5).

Once we obtain $X$ and $y$, we implement KIMI, starting from calculating the knockoff statistics $V_i$ (4) for each $k$-mer, followed by estimating the Knockoffs threshold $c_{\tau}$, and selecting $k$-mers with knockoff statistics larger than $c_{\tau}$. Then we calculate the FDP as in equation (1). The corresponding TDP is also calculated using equation (2).

The above process is repeated 100 times and the means of FDP and TDP are used as estimates of FDR and power, respectively.

Several parameters may impact the overall performance of $k$-mer selection: 1) the number of relevant $k$-mers $s$, 2) the signal estimated precision matrix is calculated by inverting the normalized sample covariance matrix. The original knockoff frequencies are normalized corresponding to the normalized covariance matrix. For easy presentation, we slightly abuse the notation and still use $x_i$'s to denote the original variables and $\hat{x}_i$'s to denote their knockoff copies.
amplitude $a$, 3) the number of contigs $n$, 4) the contig length $l$, and 5) the target FDR $p$. We will keep other parameters unchanged while studying the relationship between one particular parameter and power/FDR. Notice that the choice of $k$-mer size of interest is usually at the user’s discretion depending on the scientific problem as long as $n > p$ is not violated, while we acknowledge that larger $k$ will likely result in higher collinearity of the variables which may make the sample covariance matrix numerically singular, as shown in Section 3.1.

### 2.4.2 Simulation procedures for FDR control using the BH procedure

The BH procedure based on $p$-values has been widely used for variable selection with FDR control. To adapt to our problem, three potential approaches for calculating the $p$-values are applicable. The first is to compute the $p$-values from joint logistic regression using R function “gim”. We do not consider regularized logistic regression because no available package could output $p$-values in this scenario to the best of our knowledge. Specifically, we carry out a logistic regression on data $(X, y)$ to obtain the $p$-value for each $k$-mer indicating the statistical significance against the null hypothesis that the regression coefficient is 0 conditional on all other $k$-mers. Therefore, a very small $p$-value suggests that the corresponding $k$-mer is relevant. Here the data $(X, y)$ are the same as in Section 2.4.1.

The second is to compute the $p$-value for each column independently, which we refer to as marginal $p$-values. Note that these marginal $p$-values are for testing the marginal independence of each $k$-mer with the response, which is generally different from our goal of testing the conditional independence of one $k$-mer with the response given all others. Nevertheless, this alternative method has been very popularly used especially in high dimensions when joint regression is difficult. We use the Wilcoxon-Mann-Whitney (WMW) test (Mann and Whitney, 1947) to compare the transformed $k$-mer frequency for two types of contigs in this scenario.

Post-selection inference provides another perspective of data-driven variable selection by first selecting potentially relevant predictor variables and then carrying out statistical inference on those selected variables (Berk et al., 2013). We use the R package “selectiveInference” to calculate $p$-values of conditional (post-selection) hypothesis tests for lasso (Tibshirani et al., 2016) as the third option for producing $p$-values.

After obtaining the joint $p$-values from logistic regression, the independent $p$-values from the WMW test or the conditional post-selection $p$-values for lasso for each of the $k$-mers, we use the BH procedure to adjust the $p$-values at target FDR $p = 0.2$. Thus, the selected $k$-mers and corresponding FDR and power can be computed from 100 rounds of repetitions.

We also compare KIMI with the $q$-value method that has been commonly used for FDR control (Storey, 2002). The details are the same as those for the BH procedure except for the $p$-value adjustment and are thus omitted.

The existing literature only shows that the BH and $q$-value approaches work well under independence or certain types of dependence assumptions of $p$-values (Gavrilov et al., 2009; Reiner-Benaim, 2007; Storey et al., 2003; Jung, 2005). Since the $k$-mer frequencies depend strongly on each other, these approaches are not really applicable in such situations. We acknowledge that the above mentioned $p$-values are not ideal. We are not aware of any valid $p$-value calculating methods that completely adapt to the BH and $q$-value approaches for $k$-mer selection to the best of our knowledge.

### 2.5 Application in viral contig identification by identifying relevant $k$-mers

We next present the usefulness of KIMI in real data analysis, applying KIMI on real datasets used in VirFinder (Ren et al., 2017) for viral motif identification. Due to the lack of ground truth, we will evaluate the performance of KIMI by measuring the prediction accuracy of contig labels based on only the $k$-mers selected by KIMI. In Ren et al. (2017), all but one $k$-mer were used in lasso-penalized logistic regression for viral contig prediction. Although lasso can produce a sparse model with a set of selected $k$-mers, there is no guarantee on FDR control.

For each contig length $l$, we use the same training and testing sets as in Ren et al. (2017). Next, we fix the $k$-mer size $k$, count the number of occurrences of each $k$-mer, and normalize the count into frequency for each contig in the training set. The $k$-mer with highest $p$-value of WMW test comparing the distributions of $k$-mer frequencies in viral and bacterial contigs is then dropped. We further normalize the design matrix following the procedure discussed at the end of Section 2.3 and denote the resulting data as $(X, y)$.

Then we apply KIMI to select relevant $k$-mers $\hat{S}$ based on $(X, y)$ from the training set according to details presented in Section 2.3. We set the target FDR as $p = 0.2$. In addition to the current framework of KIMI which drops one $k$-mer to avoid perfect collinearity, we also select relevant $k$-mers using knockoff inference on the log-ratio transformed data matrix and compare the prediction accuracy with that of KIMI. The details are shown in the Supplementary Materials.

KIMI serves as a screening step for selecting relevant $k$-mers that will be useful for viral contig prediction. For prediction purpose, we only include $k$-mers in $\hat{S}$ as predictors. Using these selected features, a logistic regression model with $L_1$ penalty, which is identical to the one VirFinder used, is trained using the training data, and is then applied to the testing data to predict the probability of each contig in the testing set coming from the viral group. A threshold is then chosen, and contigs with predicted probabilities above it are classified to the viral group. True positive rate (TPR) and false positive rate (FPR) vary with the threshold. To better compare the prediction performances based on KIMI and VirFinder, we plot the receiver operating characteristic (ROC) curve by varying the threshold. The area under the ROC curve (AUC) is calculated as the measure of accuracy. For VirFinder, AUC values based on 30 rounds of bootstrap of the prediction scores in the testing set are presented. For KIMI, we generate the knockoff $k$-mer frequency matrix $\mathbf{X}$ for 30 times and calculate the AUC value for each time. We compare the AUC values based on KIMI and those presented in VirFinder.

### 3 Results

#### 3.1 The relationship between FDR/power and various parameters in simulation studies

In this section, we present simulation results answering the following questions. First, will KIMI control FDR and yield high power simultaneously for different simulated scenarios? Second, how do FDR and power change with different parameters involved in simulations? The results shown below are based on the selection of 4-mers based on simulation procedures described in Section 2.4.1 with the Markov order $r$ specified as 3. Hence, the dimensionality is $p = 255$, after dropping one $k$-mer. Although KIMI only requires...
$n > p$, a larger $k$ will generally increase the collinearity between $k$-mer frequencies, as shown in Fig. 1. We focus on the maximum correlation because it is the key factor dominating the numerical behavior of computing the knockoff statistics by penalized logistic regression and constructing the knockoff variables by inverting the covariance matrix as the precision matrix.

To study the relationship between FDR/power and a parameter, we vary a particular parameter while keep all others fixed. Unless declared, we set the following default values for the parameters: 1) the number of relevant $k$-mers $\kappa = 20$, 2) the signal amplitude $a = 0.2$, 3) the number of contigs $n = 3000$, 4) the contig length $L = 1$kbp and 5) the target FDR $\rho = 0.2$.

We start by exploring how the power and FDR change with the number of relevant $k$-mers $\kappa$. All other parameters are fixed at their default values. The power and FDR for different values of $\kappa$ are shown in Fig. 2 a. As shown therein, the power decreases with $\kappa$, which is reasonable given that selecting all relevant $k$-mers becomes harder when the target set grows larger. The power remains generally high, i.e., around or above 0.9 though. In addition to achieving high power, we also observe that the FDR is mostly controlled.

Next, we explore the relationship between FDR/power and the signal amplitude $a$. It is shown in Fig. 2 b that there is an obvious increasing trend for the power with $a$. The power at $a = 0.1$ is around 0.25 and increase steeply to above 0.9 at $a = 0.2$. The growth then slow down until the power reaches almost 1 at $a = 0.3$. The variation of power also decreases with $a$, showing larger signal amplitude helps stabilize the performance of KIMI. Meanwhile, the FDR keeps under control in all scenarios.

The relationship between FDR/power and the number of contigs in one round is presented in Fig. 2 c. Generally the trend of power is similar to what is shown in Fig. 2 b, which is reasonable given larger sample size also strengthens the signal. The FDR also slightly increases with $n$ but remains safely under controlled.

We show in Fig. 2 d that the power also rises with the contig length $L$, which can be explained as the $k$-mer frequency in longer contigs tend to become stationary. The FDR is again strictly under controlled for all values of $L$.

Among all 2000 rounds of simulations of which results are shown in Fig. 2, the relevant $k$-mer is dropped in only 5 rounds. In general, the chance of dropping a relevant $k$-mer is very rare in simulations.

Due to page limitations, we present the effect of the target FDR $\rho$ on the power and FDR as Fig. S1 in the Supplementary Materials.
3.2 The BH procedure and the $q$-value method produce very low power compared to KIMI.

We next investigate the FDR and power of the BH procedure and the $q$-value method using the similar simulation procedures as for KIMI. All default parameters are used as in Section 3.1 for KIMI, the BH procedure and the $q$-value method.

The results for comparing the performances of KIMI, the BH procedure, and the $q$-value method are summarized in Fig. 3. In contrast to KIMI that simultaneously achieves controlled FDR and high power at around 0.9, both the BH procedure and the $q$-value method regardless of using joint, marginal or post selection $p$-values failed to produce noticeable power, i.e., the power remains almost 0, making these methods meaningless despite that the FDR is controlled. In addition, the variation of FDR from both methods is larger than that of KIMI. The failure of both methods using $p$-values from joint logistic regression and post selection inference for lasso is possibly due to the very high collinearity of $k$-mer frequencies, and the marginal $p$-values are also unsurprisingly useless since they are used to test the marginal independence between each $k$-mer and the response instead of the conditional independence. Notice that the "glm" R package by default uses Wald tests to compute $q$-values from the joint logistic regression. The results based on the likelihood ratio test (LRT) are similar and are presented as Fig. S2 in the Supplementary Materials. These results show that neither the BH procedure nor the $q$-value method is applicable in our problem, highlighting the necessity and usefulness of KIMI for motif discovery.

3.3 Prediction accuracy of viral contigs based on relevant $k$-mers selected by KIMI is higher than that of VirFinder.

To see the applications of KIMI for motif identification in real data analysis, we test KIMI on contigs sampled from real viral and bacterial genomes, as discussed in Section 2.5. We fix the $k$-mer size of interest at 3 as a simple example. Our objective is to carry out a prescreening of $3$-mers by KIMI, train a classification model based on logistic regression and lasso regularization on those selected relevant $3$-mers, and predict viral contigs on the independent testing set with a focus on selecting $3$-mers with controlled FDR.

We next investigate the prediction accuracy of viral contigs in the testing data using the selected $3$-mers. The AUC values measuring the prediction accuracy on the testing set for different contig lengths by KIMI and VirFinder are shown in Fig. 4a showing that the AUC values using the $3$-mers selected by KIMI are consistently higher than those of VirFinder for all contig lengths. For example, although longer contigs tend to have higher prediction accuracy in general, the prediction accuracy based on KIMI selected $k$-mers at contig length 3kbp (5kbp) is higher than that based on VirFinder using all $k$-mers at 5kbp (10kbp), respectively. The results show the superiority of KIMI selecting relevant $k$-mers since we are using the same feature type ($k$-mer frequency) and prediction model (logistic regression with lasso regression). The improved prediction accuracy provides evidence on the relevance of $k$-mers selected by KIMI.

We also show the fractions of selected relevant $k$-mers among all $k$-mers in Fig. 4b. Around 0.6-0.75 of all the $k$-mers are selected by KIMI for different $k$-mer lengths, indicating that majority of the $k$-mers may have different frequency compositions in viral and bacterial contigs. The large proportion of selected $k$-mers may be due to the high correlation between $k$-mer frequencies or diverse compositions of $k$-mer abundance in viral and bacterial genomes. We also present the consistency of $k$-mers selected by KIMI in each round since the construction of knockoff variables involves random sampling. As shown in Fig. 4c, among the $k$-mers that have been selected by KIMI in at least 1 round, around 80% of those have been consistently selected in all 30 rounds, showing that KIMI is highly stable for $k$-mer selection in this real dataset. The fraction will increase to around 85% to 90% if we lower the threshold to 24 rounds (80% of all rounds).

Lin et al. (2014) developed a method to predict responses based on composition data using log-ratio of the relative frequencies. We extend KIMI to this model and investigate the prediction accuracy of viral contigs with the same data set based on KIMI selected variables. We show that the prediction accuracy using the log-ratio is lower than that presented in this paper. Details of the results are given as Fig. S3.
4 Discussion and conclusions

Identification of motifs relevant to specific types of molecular sequences is of paramount importance for understanding the composition of genetic materials of particular species, accurately classifying metagenomic reads, and capturing the underlying mechanisms of how different types of molecules evolve, interact with other molecules, impact public health, etc. Although many existing methods may be applied for identifying k-mers, or motifs, it is still crucial to develop methods that can enhance the reproducibility, or equivalently, control the FDR and no such methods are yet available yet to the best of our knowledge. We develop KIMI, a wrapped framework based on knockoff inference for identifying k-mers relevant to binarily labeled contigs using their occurrence frequency with guaranteed FDR control. Several key findings were shown in this paper. First, we adapted the Model-X knockoffs framework to the motif identification problem despite the difficulty from high collinearity among k-mer frequencies. Second, we showed through simulation studies that KIMI could simultaneously yield a high power and controlled FDR, while the popularly used BH procedure and γ-value method for FDR control suffer from producing very low power. Third, we presented that KIMI could be reliably used in real data analysis by showing that the prediction accuracy of contig labels could be increased with only those k-mers selected by KIMI used, compared to ViFinder. The results are consistent for all contig lengths. We expect that KIMI can be generalized for any binary types of contigs assembled from real metagenomic data and output k-mers under guaranteed target FDR level.

In spite of the key findings, KIMI also has some limitations. First, KIMI assumes that the sample size is greater than the k-mer size, which may be violated if the sample size is not large enough and large k is being investigated. Second, even if n > p is guaranteed, the high collinearity caused by large k may make the sample covariance matrix numerically singular, prompting challenges on the construction of knockoff k-mer frequencies. Finally, KIMI currently deals with two types of molecules. In many real word problems, there are many different types of molecules and it is important to extend our framework to multiple types of molecules. This is a topic for future research.

Acknowledgements

We thank Drs. Michael S. Waterman, Jinchu Lv, and Mr. Zifan Zhu at the University of Southern California for helpful discussions related to the project.

Funding

This research was partially supported by US National Institutes of Health (NIH) [R01GM120624, 1R01GM131407].

References


Reiner-Benaim, A. (2007). Fdr control by the bh procedure for two-sided cor-
related tests with implications to gene expression data analysis. Biometrical 
Journal, 49(1), 107–126.
Reiner, G., Schbath, S., and Waterman, M. S. (2000). Probabilistic and sta-
tistical properties of words: an overview. Journal of Computational Biology, 
7(1-2), 1–46.
a novel k-mer based tool for identifying viral sequences from assembled 
Ren, J., Song, K., Deng, C., Ahlgren, N. A., Fuhrman, J. A., Li, Y., Xie, X., 
Poglin, R., and Sun, F. (2020). Identifying viruses from metagenomic data 
Ricós, C., Iglesias, N., García-Lario, J.-V., Simón, M., Cava, F., Hernán-
(2007). Within-subject biological variation in disease: collected data and 
mining viral signal from microbial genomic data. PeerJ, 3, e985.
Sinha, R., Abu-Alli, G., Vogtmann, E., Fodor, A. A., Ren, B., Amir, A., Schw-
ager, E., Crabtree, J., Ma, S., Abnet, C. C., et al. (2017). Assessment of 
variation in microbial community amplicon sequencing by the Microbiome 
Quality Control (MBQC) project consortium. Nature Biotechnology, 35(11), 
1077.
P. R., Armita, J. M., and Herndl, G. J. (2006). Microbial diversity in the deep 
sea and the underexplored “rare biosphere”. Proceedings of the National 
Academy of Sciences, 103(32), 12115–12120.
of the Royal Statistical Society: Series B (Statistical Methodology), 64(3), 
479–498.
Storey, J. D. et al. (2003). The positive false discovery rate: a bayesian 
Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016). Exact 
potential inference for sequential regression procedures. Journal of 
the American Statistical Association, 111(514), 600–620.
Waterman, M. S. (1995). Introduction to computational biology: maps, 
sequences and genomes. CRC Press.