Published in final edited form as:

Stat Biosci. 2018 April ; 10(1): 41–58. doi:10.1007/s12561-017-9187-y.

A two-stage hidden Markov model design for biomarker detection, with application to microbiome research

Yi-Hui Zhou^{#1,*}, Paul Brooks², and Xiaoshan Wang^{#3}

¹Department of Biological Sciences, Bioinformatics Research Center, North Carolina State University, North Carolina, United States of America

²Department of Statistical Sciences and Operations Research and Department of Supply Chain Management and Analytics, Virginia Commonwealth University, Virginia, United States of America

³IMEDACS, LLC, United States of America, xsswang@gmail.com

These authors contributed equally to this work.

Abstract

It has been recognized that for appropriately ordered data, hidden Markov models (HMM) with local false discovery rate (FDR) control can increase the power to detect significant associations. For many high-throughput technologies, the cost still limits their application. Two-stage designs are attractive, in which a set of interesting features or biomarkers is identified in a first stage, and then followed up in a second stage. However, to our knowledge no two-stage FDR control with HMMs has been developed. In this paper, we study an efficient HMM-FDR based two-stage design, using a simple integrated analysis procedure across the stages. Numeric studies show its excellent performance when compared to available methods. A power analysis method is also proposed. We use examples from microbiome data to illustrate the methods.

Keywords

Biomarker; False discovery rates; Hidden Markov model; Metagenomics; Metatranscriptomics; PCR

INTRODUCTION

High-throughput genomic and transcriptomic sequencing technologies promise to yield great advances in biomedicine, but the enormous amount of data produced raises statistical challenges. In sequence-based research, due to the cost of processing the large number of sequences, the number of subjects sampled for a given project is still often on the order of tens or hundreds—much lower than the number of genetic features (or species) being compared between groups of subjects. The presence of a large number of features (such as genes, or bacterial taxa) being compared per subject is a familiar “large p small n ” statistical problem that can increase false discovery rates as well as decrease study power.

* yihui_zhou@ncsu.edu (YHZ) .

In biomarker discovery research, this problem is often addressed by using two-stage (or in general, multistage) designs [1]. The first stage serves to assay a large number of different features or biomarkers (such as SNPs, genes, or bacterial taxa), obtaining preliminary estimates of differences in biomarkers' presence or levels between two or more groups. The top-ranking hits from the first stage are tested in a different set of subjects at stage two, and data are analyzed jointly over both stages together [1, 2]. To our knowledge, such two-stage designs have not yet been applied in settings where hidden Markov models (HMM) are appropriate.

The two-stage study design we describe should not be confused with two-stage covariate-based sampling or multistage survey sampling, approaches that are well recognized in microbial ecologic research [2]. The goal of the latter approaches is purposeful sample selection such that results of assays conducted at stage one provide information about the most informative locations or subjects to (re)sample at the second stage [2–4]. In contrast, the goal of the two-stage approach we investigate here is to select biomarkers—not samples—for further testing. More broadly the goal is to optimize the balance of false positive and false negative discovery rates across a large number of biomarkers [1]. Although single-stage designs have higher power than two-stage designs for a fixed total sample size if all features are examined, for a fixed overall cost, two-stage procedures are generally more powerful than single-stage designs [5–7]. The novelty in this paper lies in the approach to two-stage (or multi-stage) testing, assuming the data follows the HMM structure.

Much of the methodologic development of optimal two-stage study designs has been performed in the context of genomewide association studies [8]. Although SNP markers have correlation structure, it is generally thought that regional association is due to one or more causal SNPs [9], with the remaining correlations reflecting linkage disequilibrium. In contrast, microbial metagenomic sequencing identifies large classes of taxa that may collectively show association with an outcome or phenotype. We consider a specific motivating context in which, in stage 1, data on a very large number of potential disease-associated candidates are generated through an expensive technique. A technique suitable for a follow-up sample is used in the second stage, targeting specific top-ranked candidates from stage one (e.g. in microbiome studies, quantitative PCR or RT-PCR for selected taxa). Here, “disease-associated” differences could substitute equally for differences across sampling sites, across time points, or across groups defined by any other feature.

It would be valid to treat the first stage as a pilot arm, i.e. pilot design, in which the first stage data serves only for screening, and to draw conclusions based only on the second-stage data. This wastes useful information from the first stage, however, and integrated analysis is almost always more statistically powerful than separate analyses [5, 6, 10]. In previous applications, joint analysis was performed by combining the data from each stage [1]. This approach requires similarities of the assay platform and effect measures, however, that do not apply when two completely different assay techniques are used, such as sequencing followed by PCR.

We propose a novel hidden Markov model based two-stage analysis that addresses both of these challenges. To address the dissimilarity of measures at each stage, we perform joint

two-stage hypothesis testing by averaging the local index of significance from each stage. We show that, with modest regularity conditions, the procedure achieves the goal of controlling the FDR. Examples of microbial sequence data are used for illustration. Through simulation studies we demonstrate that the procedure gives a lower false negative rate (FNR) than the method proposed by Zehetmayer et al [11] and p value combination methods [12]. We develop algorithms and software that can be used for sample size calculations to aid in planning such two-stage studies.

Methods

We use the term “biomarker discovery” loosely to denote comparing two (or more) groups in a search for differences in any type of biological or biochemical feature when many such features are compared simultaneously. The study design we propose is general and applies to any context in which a large number of hypotheses are being tested with practical constraints on sample size or cost.

Consider an experiment comparing two groups or treatments. Without loss of generality, we apply a Z -score transformation to the differential test statistics. Thus our method applies equally well to a quantitative/continuous experimental variable, but we focus here on the two-group problem for specificity. Denote the hidden status of feature j by θ_j , $j = 1, \dots, m$ where $\theta_j = 1$ if feature j differs between the diseased and healthy groups (i.e. non-null) and 0 otherwise (null). We assume that $Z_j | \theta_j \sim f_{\theta_j}$, $\theta_j \in \{0, 1\}$ and that Z -scores are independent of each other conditional on the unobserved status θ_j . The overall probability that a random feature is null is π_0 . Typically $f_0(Z)$ is the $N(0, 1)$ density, although in our implementation we assume only normality, and the mean and variance are fitted from the data. In our modeling (described below), $f_1(Z)$ is a finite mixture of normal densities, where the mixtures components are also fitted from the data.

Recently, Sun and Cai introduced a Hidden Markov Model (HMM) based approach that incorporates dependence structure of successive θ_j values into the multiple testing framework [13]. This HMM based method is optimal when the HMM parameters are known. When the parameters are unknown, an EM algorithm is proposed to obtain consistent estimates of parameters, and the procedure with plug-in the estimates is asymptotically optimal. Specific details of the estimation procedure are provided below, but the key elements are estimation of f_1 and a transition matrix A_0 for θ . Following Sun and Cai[13], we consider a normal mixture for f_1 , with a number of specified mixture components. The primary novelty of Sun and Cai [13] was to introduce a likelihood and modeling framework to provide a local index of significance (local FDR) for each feature, thus enabling overall FDR control with high power.

Despite the interest in the HMM-based FDR control, there are relatively few data types that clearly natively follow an HMM structure, as the correlation structure of [13] is assumed to follow only from the hidden status θ_j . We propose that microbial metagenomic datasets, where sequenced instances represented taxa abundances, may be reasonably modeled using an HMM. The rationale is that large sets of taxa, organized using high order groupings, tend to be null or alternative together. Ordering the taxa by standard taxonomy (hierarchically,

from phyla to genera) will produce “transitions” of null/alternative status, in a manner that mimics a Markov process. Although the taxa properties are not truly random, the success of the method can be judged in terms of number of significant findings, given that the FDR is well-controlled.

As an extreme instance of grouping, and not assuming knowledge of the true taxonomic structure, we might also consider performing hierarchical clustering of the features. Such clustering may produce an artificial ordering that may be partly driven by non-specific correlation structures, irrespective of the hidden status. Following such clustering, serial correlation in the Z_j values may be present, although such correlation is typically much smaller than the correlation of clustered features. Thus we describe in the supplement a decorrelation procedure that may be used with clustered data. In our examples, the clustering order algorithm follows the convention that “tighter” subclusters (i.e. those with smaller average distance metrics for the clustered elements) are placed first in the ordered sequence.

Determining the LIS thresholds and power

In order to have the best power to detect the taxa which are truly differential between two groups while the overall FDR is controlled, we need to find a threshold value for Stage 1 to ensure that a large proportion of truly differential taxa is kept for Stage 2 retesting, and to control the overall FDR after both stages. We assume m tests, and we define the LIS for test/feature j as the local FDR [13, 20], $LIS_j = P(H_0 | \text{data}) = \pi_0 f_0 / f$, where $f = \pi_0 f_0 + (1 - \pi_0) f_1$. I.e. the LIS_j is the probability of the null for feature j , given the data, using plug-in estimates of f_0 , f_1 , π_0 , and A_0 . Let $\gamma_j^{(i)}$ be the LIS of the j th feature in the i th stage, and for any collection of features Ω , the FDR is the mean of the LISs, or $\gamma^{(i)} = \sum_{j \in \Omega} \gamma_j^{(i)} / |\Omega|$. For $\gamma_j^{(i)}$ ordered from smallest to largest, $\gamma^{(i)}$ thus serves as a q -value [21], and for any intended FDR α , we can find the threshold for Z such that $\gamma^{(i)} = \alpha$.

For the special case in which the alternative f_1 is a single $N(\mu_1, 1)$ density, and for two-sided testing, we can determine the power as follows. By a simple calculation and in general,

$$\frac{\gamma_j^{(i)} - \pi_0}{1 - \pi_0} = \frac{f_0(z_j)}{f_1(z_j)}$$

Suppose j is indexed such that the $\gamma_j^{(i)}$ are increasing, and a critical value $c^{(i)}$ determines Ω , corresponding to Z s which are more extreme than $c^{(i)}$. For this setup, $c^{(i)}$ must satisfy

$$\frac{\gamma^{(i)} - \pi_0}{1 - \pi_0} = \frac{2\Phi(-c^{(i)})}{1 - \Phi(c^{(i)} - \mu_1) + \Phi(-c^{(i)} - \mu_1)}. \quad (1)$$

where Φ is the cumulative distribution function of a normal distribution.. The right hand side of (1) is $\frac{F_0(z)}{F_1(z)}$. Since the normal distribution has the monotone likelihood ratio property [22],

the cdf ratio $\frac{F_0(z)}{F_1(z)}$ is an increasing function of z for $\mu_1 > 0$, such that a unique solution of $c^{(j)}$ exists. Once $c^{(j)}$ is determined, the corresponding power becomes $S^{(j)} = 1 - \Phi(c^{(j)} - \mu_1) + \Phi(-c^{(j)} - \mu_1)$.

An HMM simulation of single-stage HMM modeling

We consider a simulation experiment, with sample size $n=100$, $m=1000$ features and 20 true clusters of taxa in a hypothetical metagenomics experiment with species abundance compared to a dichotomous phenotype. Specifics of the simulation are provided in Supplementary Information, including the simulation of realistic discrete abundance outcomes and correlation-based hierarchical clustering of taxa (i.e., naïve to the true cluster membership). If desired, a decorrelation step is proposed (see Supplementary material) to ensure that the Z values do not violate the HMM assumptions, although in practice the results before/after decorrelation are similar. The simulations were performed to carefully mimic the properties of microbial taxa (Supplement), and for some taxa the count data can include many zeroes. Accordingly, the Fisher z -transformation was applied to the Pearson correlation coefficient between the dichotomous experimental variable and the data from each taxon. The Fisher z is remarkably robust. In the Supplement we demonstrate via simulation for a relatively small sample size ($n=50$) and a hypothetical feature with very low mean (Poisson with mean 0.3) that z is approximately normally distributed with variance near 1, either under the null or a specified alternative.

Estimation and step-up FDR control using the local index of significance are described in detail further below. In the simulations, one of the true clusters was simulated as “alternative” for all taxa in a single true cluster, with the strength of the alternative determined by a parameter β . Results in terms of true FDR control and number of identified taxa for this single-stage HMM modelling are shown in Supplementary Figure 1, and compared to standard Benjamini-Hochberg FDR control. We point out that, while Benjamini-Hochberg FDR control is often modestly conservative, the Benjamini-Yekutieli procedure [14] has weaker dependency assumptions and could also be used, although it is even more conservative than Benjamini-Hochberg and would thus have lower power.

The results of the simulations show that the HMM modeling approach achieves appropriate FDR control, while identifying, as expected, more significant taxa than standard Benjamini-Hochberg. The FDR control (Supplementary Figure 2) follows a familiar pattern in which, for small effect sizes, the chance that *any* rejections occur is small for either procedure. However, for large effect size, rejections are almost certain to occur, and the realized proportion of false discoveries is very near the target = 0.05. In addition, for each simulation the HMM model was fit using randomly permuted Z values, showing reasonable control of the FDR and similar number of rejected taxa as the Benjamini-Hochberg approach. The results indicate that the HMM approach, even in the case where the hidden structure is not apparent to the researcher, essentially does no harm.

Two-stage modeling—To introduce the two-stage modeling, we begin with notation covering both the first and second stages. Suppose we have a total of m tests, where m_0 of

them are null (no association with the external variable), and m_1 are non-null (alternativ. In a two-stage design, we select a subset of biomarkers based on the first stage data, and collect data at the second stage to perform statistical tests only on this subset. Table 1 summarizes the possible outcomes from a multiple testing framework.

The overall FDR is defined as

$$FDR = E\left(\frac{V^{(1)} + V^{(2)}}{R^{(1)} + R^{(2)}} \mid R^{(1)} + R^{(2)} > 0\right) P\left(R^{(1)} + R^{(2)} > 0\right)$$

and the overall false non-discovery rate FNR is defined as

$$FNR = E\left(\frac{T^{(1)} + T^{(2)}}{A^{(1)} + A^{(2)}} \mid A^{(1)} + A^{(2)} > 0\right) P\left(A^{(1)} + A^{(2)} > 0\right).$$

The validity of a multiple testing procedure is determined by its capability to control the FDR at a predefined level. Among all such procedures that can achieve FDR control, the optimal one gives the smallest FNR. In this study, we also compare possible choices which are available for analyzing data from two-stage studies, to provide practical guidance on choosing an appropriate method.

Here, we extend the use of the HMM model to control FDR and optimize the FNR in two-stage designs. In a two-stage design, we define two thresholds γ_0 and γ_1 for first-stage screening. We declare $h_0 = \max\left\{h_0: \frac{1}{h_0} \sum_{j=1}^{h_0} LIS_{(j)}^{(1)} \leq \gamma_0\right\}$ taxa significantly differential. In the case there are no significant differential taxa, we set $h_0 = 0$. We select all $h_1 = \max\left\{h_1: \frac{1}{h_1 - h_0} \sum_{j=h_0}^{h_1} LIS_{(j)}^{(1)} \leq \gamma_1 - \gamma_0\right\}$ taxa for re-testing in the second stage. For all others, their null hypotheses are accepted. Selecting taxa with small LISs follows from the fact that a small LIS indicates a small probability to be null for its corresponding test. We defer the discussion on the choice of the first stage threshold to the next section.

With the second stage data, we run the EM algorithm to obtain a new set of HMM parameters of a normal mixture model. $LIS_j^{(2)}$ is estimated by a forward-backward procedure with the estimated second-stage non-null distribution $f_1^{(2)}$, the estimated first-stage null distribution $f_0^{(1)}$, and the first-stage transition probability $A_0^{(1)}$. The reason to use $A_0^{(1)}$, instead of estimating a new $A_0^{(2)}$, is that $A_0^{(1)}$ is estimated from the data including all taxa, and thus captures more accurate information on the transition probabilities.

Another subtle issue is due to degeneracy caused by unboundedness of parameters in the Gaussian mixture, which is more likely to occur for the second stage data, since the proportion of null hypotheses is often small after the first stage screening. We propose to apply constrained EM [15] by placing a lower limit to the ratio of standard deviation

between the non-null and the null, i.e. $\min\left(\frac{\sigma_1}{\sigma_2}\right) \geq c > 0$. Simulations indicates that such a simple constraint greatly improve the estimation procedure, by providing consistent estimates and reducing the running time of EM iteration. Analysis of datasets with up to 10,000 features typically requires only seconds or minutes on a standard desktop computer.

By the definition of LIS, the explicit form to describe the density of LIS is difficult to derive because the density ratio involves a mixture of normal distributions. An enormous amount of research has been done to model the distribution of LIS, (a local FDR), through using different density estimators, e.g, the exponential family density [16], Bernstein density [17], modified Grenander density [18], Parzen-kernel density [19] and many others. Applying such estimating methods would further complicate our situation to combine information from both stages.

To simplify the estimation and facilitate combined tests, we assume $LIS_j^{(1)}$ and $LIS_j^{(2)}$ of taxa j follow a uniform distribution, and are independent with each other (although in practice they will typically be positively correlated). We use this assumption because it is typically conservative (for small ϵ below), and greatly simplifies the process. With the “assumption,” we consider a mean of the two LISs:

$$mLIS_j = \frac{LIS_j^{(1)} + LIS_j^{(2)}}{2}, \quad (2)$$

We then rank $mLIS$ s, $mLIS_{(1)}, \dots, mLIS_{(h)}$, and find a maximum l such that

$$l = \max \left\{ l: \frac{1}{l} \sum_{j=1}^l mLIS_{(j)} \leq \alpha \right\} \quad (3)$$

and then declare all l tests are significant. We refer to this approach as mHMM-FDR. Under the conservative [20] assumption, we have for small ϵ_j

$$Pr \left(\frac{LIS_j^{(1)}}{2} + \frac{LIS_j^{(2)}}{2} \leq \epsilon_j \right) \leq \epsilon_j.$$

The result implies that $mLIS_j$ can be conservatively used like a real LIS of a single stage design. It follows that

$$\frac{1}{l} \sum_{j=1}^l mLIS_{(j)} \leq \alpha.$$

Result 1: Consider a two-stage design for multiple tests on a normal mixture with component-wise status defined by an HMM. Define the test statistic $mLIS_j$ which is a mean

of LISs from each stage. Let $mLIS_{(1)}, \dots, mLIS_{(h)}$ be the ranked LIS values and $H_1, \dots, H_h$ the corresponding hypotheses. Then the testing procedure (2) controls FDR at α asymptotically.

We consider a limited resource for Stage 2, which means we typically select a small proportion p of m into the second stage. Then, we are constrained by

$$2\pi_0\Phi(-c^{(1)}) + (1 - \pi_0)S^{(i)} \leq p \quad (4)$$

To determine the Stage 1 threshold $\gamma^{(1)}$, we can use (4) to find $c^{(1)}$, and substitute it into (1) to obtain an estimate of $\gamma^{(1)}$. Power calculation is straightforward. We use (1) to find $c^{(2)}$, and the second stage power is $S^{(2)} = 1 - \Phi(c^{(2)} - \mu_1) + \Phi(-c^{(2)} - \mu_1)$. Since the two stages are independent with each other, the overall power is the product of powers from each stage, i.e. $S = S^{(1)}S^{(2)}$.

Then, with the known total number of taxa $m, \mu_1^{(i)}$ for each stage, the proportion of true null π_0 , the total of sample size n per group, the pre-defined overall FDR level α and the overall power S , we can apply a simple grid search to find an optimal allocation of sample size which gives the best power.

Simulations to evaluate control of FDR and statistical power

We generated 1,000 data sets for each scenario described as follows. To mimic real data, we generate the total number of biomarkers tested $m = 500$, and 1000. We compare the proposed mHMM-FDR method with Zehetmayer's approach [1], which we refer to as the Z approach, and with Fisher's combined test as proposed by Sarkar et al [11]. As a reference comparison, we can assume all tests are retested in the second stage, and the data can be combined directly, so that we can apply a HMM-FDR on the full data (full HMM-FDR). A single stage Benjamini-Hochberg procedure is also included.

Each Z_j was generated from a normal distribution, conditional on the unobserved state θ_j . For each m , a predefined total observation N , and an effect size δ , we find an optimal design based on Z approach. Data is generated based on the optimal design, and used by both the Z approach and the proposed new method. We vary a_{11} of the transition probability matrix for θ ,

$$A_0 = \begin{pmatrix} 0.95 & 0.05 \\ 1 - a_{11} & a_{11} \end{pmatrix}$$

to vary dependence structures among non-null cases. A large a_{11} indicates that nonnull cases are more likely to form clusters. Each Z_j was generated from a normal distribution conditional on θ_j with mean zero under the null and mean δ under the non-null, i.e. $Z_j|\theta_j \sim (1 - \theta_j)N(0, 1) + \theta_jN(\delta, 1)$.

To demonstrate the effect of misspecification in the number of mixture components on FDR, we all simulate from a two-state HMM with $Z \sim \mathcal{N}(0,1)$ under the null, while under the non-null, Z is drawn from a three-component normal mixture $0.4\mathcal{N}(\mu, 1) + 0.3\mathcal{N}(-1,1) + 0.3\mathcal{N}(-3,1)$. Under this three-component alternative, we misspecify f_1 as following a two-component model. The transition probability matrix is taken to be $a_{00} = 0.95$ and a_{11} , where we vary a_{11} between 0.2 and 0.8. In addition, we also vary μ from -4 to -1 in increments of -0.5 .

Software implementing our proposed framework is available as an R package “TwoStage”. The software can be found at <https://github.com/xsswang/twostage>

Application to the HMP dataset

The NIH Human Microbiome Project (HMP) [23] analyzed a large cohort, revealing the extent to which individuals can differ in the microbes which inhabit the gut, skin, etc.. To illustrate our statistical method, we use the stool data, consisting of 88 females and 103 males in their first clinical visit. One objective is to identify species associated with sex, potentially reflecting the influence of hormonal differences [24]. We excluded taxa with zero variation in either females or males (typically due to zero tag counts for all samples), and 748 taxa remained in the analysis. The ordering of bacterial taxa was performed using known taxonomy (Phylum, Class, Order, Family, Genus, Fig. 1 (A)). A scatterplot of cross-taxa correlations for these data shows striking block structure (Fig. 1 (B)). Solid lines indicate family-level boundaries in the ordered taxa, showing that the correlation boundaries correspond to taxa relationships (the Bacteroidaceae family clearly shows additional substructure within). For downstream analysis, each tag count entry in the data matrix was divided by the total library size (sum of tag counts) for that sample to create a normalized dataset. In order to mimic two-stage sampling, we divided the samples into 85 for a stage 1 set (37F, 48M), and 106 for a second stage followup (51F, 55M), testing associations only with taxa meeting the stage 1 screening criteria.

For each of the stages, a z -statistic was formed by first calculating the Pearson correlation coefficient r_j for each taxon j between the library-normalized metagenomic data vs. the sex status (0=male, 1=female). Then the Fisher transformation $z_j = \frac{1}{2} \log \left(\frac{1+r_j}{1-r_j} \right)$ was calculated.

One advantage of the Fisher transformation, in comparison to other statistics, is the fact that the variance remains near 1 for entire range of possible true correlations, fitting the HMM assumptions.

The dependence among test statistics in stage 1 is the property necessitating novel approaches. As mentioned above, integrated designs generally give improved statistical power by combining information from both stages [5]. However, next generation sequencing gives taxon-specific gene counts that depend on the depth of sequencing, while qPCR relies on species-specific primers and amplification steps that can distort the absolute and relative bacterial counts. These differences may lead to heterogeneity in the magnitude of disease association with a given taxon between stage one and stage two, and thus flexible integration of evidence is necessary.

Results

Simulation studies

For comparison, we included Zehetmayer's Z method, Fisher's combination of p values, the single stage Benjamini-Hochberg (BH) approach, and the full data HMM-FDR procedure. In the simulation studies, we run each procedure on the generated data. The average FDR, FNR, and total number of positives are reported from each scenario.

Both the Z method and Fisher's combination test perform well in controlling the FDR. The proposed mHMM-FDR method maintains a good control at the nominal FDR (Figures 2 and 3). Its performance is close to the full-data HMM-FDR. It is relatively conservative with respect to FDR when non-null hypotheses are more likely to form correlated clusters, i.e. when $a_{11} > 0.5$ and the effect size is fixed at 1.5 (Top row of Figs. 2 and 3), when comparing to the full-data HMM-FDR which shows an increase of FDR with a_{11} . However, in these same scenarios, the mHMM-FDR procedure also gives a good FNR even under positive dependence of the tests ($a_{11} > 0.5$), whereas Z, Fisher's combination, and single-stage BH procedures rapidly increase FNRs when $a_{11} > 0.8$, suggesting that they have low abilities to detect non-nulls when non-nulls have strong relationship.

All methods control the FDR across a range of effect sizes. However, the mHMM consistently gives FNRs close to the full-data HMM-FDR method and has a good ability to detect non-nulls (bottom row of Figs. 2 and 3). The Z method has a slightly better FNR when the effect size is beyond 2.5.

Misspecifying the number of mixture components of f_1 does not affect mHMM's ability to control FDR. With misspecification, mHMM has FNR close to the full data HMM-FDR, but much lower than other methods when a_{11} is less than 0.7. However, it has much smaller FNR when there are strong correlations, so that it maintains a better capability to detect non-null. Both the Z approach and Fisher's combination method has a very limited ability, as demonstrated by Figs 4 and 5.

Analysis of HMP data

We applied the proposed mHMM-FDR procedure to the HMP stool Z-statistics Stage 1 as described (85 samples), for association with sex. Two components of f_1 were selected based on the Bayesian Information Criterion (BIC, Table 2). If we treat this study as the first stage of a two-stage design, 16 taxa were significant at $FDR < 0.05$ according to the mHMM-FDR procedure, and passed to the second stage retest. After the second stage (106 samples, using the metagenomics data on the 16 taxa as a stand-in for qt-PCR sampling), 15 of the 16 taxa remained significant at $FDR < 0.05$. The final 15 taxa include families Porphyromonadaceae, Bacteroidaceae, Rikenellaceae, Lachnospiraceae, Ruminococcaceae. Interestingly, for all of the significant taxa, males showed increased relative abundance compared to females (Figure 6). We note that at a higher level of evolutionary organization, there appears to be substantial overlap with the sex-biased gut microbiome findings of [24] in a mouse model of type 1 diabetes, although the authors did not find such consistent male over-abundance. Here we use the current data as a proof of principle for the general approach.

Discussion

The two stage HMM method introduced here can be used in a variety of biomarker discovery contexts for which serial correlation in null/alternative status is likely to arise naturally. The methods we have described combine the HMM modeling, which has been shown to increase power [13], with novel considerations of combining LIS information for biomarkers as features. Although analysis of two-stage microbiota data offer a good example of these techniques, we expect that similar considerations arise in a number of other contexts. Accordingly, the TwoStage software has been written in a generic fashion and can be readily applied to other data types.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Reference

1. Zehetmayer S, Bauer P, Posch M. Two-stage designs for experiments with a large number of hypotheses. *Bioinformatics*. 2005; 21:3771–3777. [PubMed: 16091414]
2. Tickle TL, Segata N, Waldron L, Weingart U, Huttenhower C. Two-stage microbial community experimental design. *ISME J*. 2013; 7:2330–9. [PubMed: 23949665]
3. Breslow NE, Cain KC. Logistic regression for two-stage case-control data. *Biometrika*. 1988; 71:11–20.
4. Haneuse S, Schildcrout J, Gillen D. A two-stage strategy to accommodate general patterns of confounding in the design of observational studies. *Biostatistics*. 2012; 13:274–88. [PubMed: 22130627]
5. Goll A, Bauer P. Two-stage designs applying methods differing in costs. *Bioinformatics*. 2007; 23:1519–26. [PubMed: 17463029]
6. Kraft P, Cox DG. Study designs for genome-wide association studies. *Adv Genet*. 2008; 60:465–504. [PubMed: 18358330]
7. Stanhope S aSkol AD. Improved minimum cost and maximum power two stage genome-wide association study designs. *PLoS One*. 2012; 7:e42367. [PubMed: 22970116]
8. Simon-Sanchez Javier, et al. Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nature genetics*. 2009; 41(12):1308–1312. [PubMed: 19915575]
9. McCarthy MI, Hirschhorn JN. Genome-wide association studies: potential next steps on a genetic journey. *Hum Mol Genet*. 2008; 15(R2):R156–65. 17.
10. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet*. 2006; 38:209–13. [PubMed: 16415888]
11. Zehetmayer S, Bauer P, Posch M. Two-stage designs for experiments with a large number of hypotheses. *Bioinformatics*. 2005; 21:3771–7. [PubMed: 16091414]
12. Sarkar S, Chen J, Guo W. Multiple Testing in a Two-Stage Adaptive Design with Combination Tests Controlling FDR. *J Am Stat Assoc*. 2013; 108:1385–1401.
13. Sun W, Tony Cai T. Large-scale multiple testing under dependence. *J R Stat Soc Ser B (Statistical Methodol)*. 2009; 71:393–424.
14. Benjamini YoavYekutieli Daniel. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*. 2001:1165–1188.
15. Hathaway RJ. A Constrained Formulation of Maximum-Likelihood Estimation for Normal Mixture Distributions. *Ann Stat*. 1985:795–800.
16. Efron B. Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis. *J Am Stat Assoc*. 2004; 99:96–104.

17. Guan Z, Wu B, Zhao H. Nonparametric estimator of false discovery rate based on Bernstein polynomials. *Stat Sin.* 2008; 18:905–923.
18. Strimmer K. A unified approach to false discovery rate estimation. *BMC Bioinformatics.* 2008; 9:303. [PubMed: 18613966]
19. Guedj M, Robin S, Celisse A, Nuel G. Kerfdr: a semi-parametric kernel-based approach to local false discovery rate estimation. *BMC Bioinformatics.* 2009; 10:84. [PubMed: 19291295]
20. Rüschendorf L. Random Variables with Maximum Sums. *Adv Appl Probab.* 1982; 14:623–632.
21. Efron B, Storey J, Tibshirani R. *Microarrays Empirical Bayes Methods, and False Discovery Rates.* 2001
22. Lehmann EL. *Testing Statistical Hypotheses.* Wiley; New York: 1986.
23. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature.* 2012; 486(7402):207–214. [PubMed: 22699609]
24. Markle Janet GM, , et al. Sex differences in the gut microbiome drive hormone-dependent regulation of autoimmunity. *Science.* 2013; 339(6123):1084–1088. [PubMed: 23328391]

Figure 1 (A).

Heatmap of the HMP data, consisting of tag counts for 748 Operational Taxonomic Units (OTUs), transformed as $\log_e(\text{count}+0.5)$, with 103 males and 88 females.

Figure 1 (B).

Heat map of sample correlations of $\log(\text{count}+0.5)$ between OTUs. Data were from a metagenomic analysis of NIH human microbiome project with 103 males and 88 females, with OTUs ordered by phylogenetic relationships. The correlations are primarily block structured, with less extreme negative correlations than positive correlations. Solid lines indicate family-level boundaries in the ordered taxa.

Figure 2.

Average empirical FDR, FNR, and Average Total Positives (ATP) for various a_{11} at fixed $\delta = 1.5$ (Row 1) and as a function of effect size δ (Row 2) for $m = 500$ at nominal FDR of 0.05. Column 1 compares the empirical FDR. Column 2 compares the empirical FNR. Column 3 compares the ATP. Methods include mHMM ($\circ$), Z approach ($\square$), Fisher's combination ($\times$), full data HMM ($\diamond$), and one-stage Benjamini-Hochberg procedure (+).

Figure 3.

Average empirical FDR, FNR, and Average Total Positives for various a_{11} at fixed $\delta = 1.5$ (Row 1) and as a function of effect size δ (Row 2) for $m = 1000$ at nominal FDR of 0.05. Column 1 compares the empirical FDR. Column 2 compares the empirical FNR. Column 3 compares the average total positives. Methods include mHMM (○), Z approach (□), Fisher's combination (×), full data HMM (◇), and one-stage Benjamini-Hochberg procedure (+).

Figure 4.

Average empirical FDR, FNR, and Average Total Positives when the number of components for nonnull is misspecified, for various a_{11} at fixed $\delta = 1.5$ (Row 1) and as a function of effect size δ (Row 2), with $m = 500$ and the nominal FDR of 0.05. Column 1 compares the empirical FDR. Column 2 compares the empirical FNR. Column 3 compares the ATP. Methods include mHMM (○), Z approach (□), Fisher's combination (×), full data HMM (◇), and one-stage Benjamini-Hochberg procedure (+).

Figure 5.

Average empirical FDR, FNR, and Average Total Positives when the number of components for nonnull is misspecified, for various a_{11} at fixed $\delta = 1.5$ (Row 1) and as a function of effect size δ (Row 2), with $m = 1000$ and the nominal FDR of 0.05. Column 1 compares the empirical FDR. Column 2 compares the empirical FNR. Column 3 compares the Average Total Positives. Methods include mHMM (O), Z approach (), Fisher's combination (x), full data HMM (◇), and one-stage Benjamini-Hochberg procedure (+).

Figure 6. Phylum, class, order, and family for the 15 significant genera identified using the two-stage sampling (genera averaged within each family)
Using the full dataset, ratios of (female mean count)/(male mean count) are shown, corrected for a slight (1.1) male:female bias among the utilized 748 genera/taxa. Bold-face indicates microbiome order/families that appeared among significant sex-based genera in Markle et al. [23].

Table 1

Testing decisions from a two-stage design.

	Declared non-significant at 1 st stage	Non decisive, retest in 2 nd stage		Declared significant at 1 st stage	Total
		Declared non-significant	Declared significant		
Null hypothesis is True	$U^{(1)}$	$U^{(2)}$	$V^{(2)}$	$V^{(1)}$	m_0
Alternative is True	$T^{(1)}$	$T^{(2)}$	$S^{(2)}$	$S^{(1)}$	m_1
Total	$A^{(1)}$	$A^{(2)}$	$R^{(2)}$	$R^{(1)}$	m

Table 2

Comparison of Gaussian mixture models for microbial metagenomic data.

f_0	f_1	Transition Matrix	Number of parameters	BIC
$N(0.66864, 0.1^2)$	$N(0.65703, 0.1^2)$	$\begin{pmatrix} 0.99 & 0.01 \\ 0.84 & 0.16 \end{pmatrix}$	7	-252.9
$N(0.66830, 0.1^2)$	$0.20N(0.65878, 0.1^2) + 0.80N(0.67568, 0.1^2)$	$\begin{pmatrix} 0.97 & 0.03 \\ 0.61 & 0.39 \end{pmatrix}$	10	-262.9
$N(0.62704, 0.1^2)$	$0.13N(0.71955, 0.1^2) + 0.59N(0.71958, 0.1^2) + 0.28N(0.71959, 0.1^2)$	$\begin{pmatrix} 0.89 & 0.11 \\ 0.13 & 0.87 \end{pmatrix}$	13	-233.5