Statistics and Its Interface Volume 0 (2014) 1–13

A Bayes testing approach to metagenomic profiling in bacteria

Bertrand Clarke*, Camilo Valdes, Adrian Dobra, and Jennifer Clarke

Using next generation sequencing (NGS) data, we use a multinomial with a Dirichlet prior to detect the presence of bacteria in a metagenomic sample via marginal Bayes testing for each bacterial strain. The NGS reads per strain are counted fractionally with each read contributing an equal amount to each strain it might represent. The threshold for detection is strain-dependent and we apply a correction for the dependence amongst the (NGS) reads by finding the knee in a curve representing a tradeoff between detecting too many strains and not enough strains. As a check, we evaluate the joint posterior probabilities for the presence of two strains of bacteria and find relatively little dependence. We apply our techniques to two data sets and compare our results with the results found by the Human Microbiome Project. We conclude with a discussion of the issues surrounding multiple corrections in a Bayes context.

AMS 2000 subject classifications: Primary 62F15, 62P10; secondary 62-07, 62F03.

Keywords and phrases: Metagenomics, Bayes testing, Bacteria, Dependence.

1. INTRODUCTION

With the growing availability of sequencing technologies the number of research contexts involving data from an unknown but possibly complex genomic source is rapidly growing. Often the source population is a mixture of multiple genomes that may be called a metagenomic population. The challenge to the statistician is to determine the composition of this population in terms of its component genomes, i.e., identify which bacterial strains or species are present and whether any may pose a risk to human health or the environment. For instance, in human health, The Human Microbiome Project (HMP) has discovered associations between microbial gut composition and obesity [9] while in agriculture, the CDC estimates that each year roughly 1 in 6 Americans (or 48 million people) gets sick, 128,000 are hospitalized, and 3,000 die of foodborne diseases [32]. Accurate and cost effective identification of bacteria at the strain level is vital for earlier detection, intervention and targeted treatment.

Detection of bacterial species has improved dramatically in recent years largely due to the development of next generation sequencing (NGS) ([26], [27]). NGS allows for detection at the whole genome level, leading to further understanding of relationships between bacterial strains and mutations specific to each strain. Recent literature indicates that whole genome NGS more accurately detects known bacterial genomes and more easily differentiates among known genomes than more traditional and targeted methods [29]. For further details about NGS data and statistical issues see [1] and [7].

The main contribution of this paper is to provide a statistical approach to the detection of bacterial genomes at the strain level from NGS metagenomic data. Our technique begins by assuming the population being sequenced is likely to contain one or more strains of bacteria along with genetic material from non-bacterial sources (e.g., human, archaea, virus). From this population a sample is taken and analyzed by whole genome NGS sequencing. There are many NGS sequencing platforms; our method assumes relatively short reads (100 bp) but other platforms can be accommodated by obvious variants of our technique. The short reads are aligned to a reference database containing M known bacterial strains. Since reads may be non-unique, we permit fractional assignment as discussed in Sec. 2. The result of this is that the data we analyze consist of the read counts for each of the M genomes in the database. Essentially, we use a Dirichlet prior on the probability of detection of each of the bacteria in the data base and regard the read counts for the genomes as multinomial. Since the Dirichlet is conjugate for the multinomial, the posterior distribution for the proportion θ_i of genome i in the population is easy to find. So, M marginal Bayes hypothesis tests can be used to decide whether or not each strain is present in the population. That is, if there are M bacterial genomes in the data base, M Bayes tests are done, one for each strain. Since the data are fixed, and hence no longer regarded as stochastic, our focus is on obtaining a single posterior density that describes the proportion of each bacterial strain in the population. This posterior is for a single $M+1$ dimensional parameter, conditioned on a single data set.

Many techniques have been developed to address detection of bacteria in metagenomic samples. As described in

*Corresponding author.

[11] these methods fall into three general categories: taxonomic mapping, composition, and whole-genome assembly. Probably the most widely used technique for taxonomic mapping is MetaPhlAn [35] which uses a carefully curated database of clade markers to identify individual species from a metagenomic sample. Although MetaPhlAn performs favorably compared to other existing methods such as PhymmBL [4], it is limited to the species for which unequivocal clade markers have been identified (roughly 25% of known species). However, currently MetaPhlAn is only used at the species level, not the strain, level. A similar but more recent method, specI [25], uses phylogenetic marker genes to identify prokaryotic species and species clusters. As the authors note, its purpose is to automate phylogenetic analysis for large-scale applications and bring more objectivity to the field of phylogeny (the same objective as PhyloPhlAn [36]). Among the composition methods the most recent is Pathoscope [11] which can be used for species or strain identification. Pathoscope works by aligning reads to genome sequences in a known database. Reads that cannot be uniquely assigned to a single genome are 'reassigned' to the single 'best' genomic source using an expectation-maximization (EM) approach based on a multinomial likelihood. Because Pathoscope uses an EM optimization, when used on the strain level it will tend to reassign non-unique reads to only one strain among many similar strains, discounting the possibility that many similar strains may be present. In this paper we do not discuss whole-genome assembly methods because, although they can be very accurate at strain identification, they require much greater sequencing coverage than is common in our applications of interest.

The method we propose here belongs to the composition class as we are not using taxonomic information nor are we attempting whole genome assembly. First, unlike other composition approaches our method provides a probabilistic assessment of the presence/absence of reference bacterial strains, and assesses the likelihood of the presence of a genome not in the current reference database. Second, as our approach does not involve an auxiliary optimization such as EM, it readily scales to thousands of reference genomes. Our reference database contains over 5,000 genomic sources while the samples may include information from tens to hundreds of strains. Third, our method focuses on strain detection which amounts to identification for genomes in our reference database. While methods for detection focus on minimizing false positives, our method is more concerned with minimizing false negatives. That is, we allow our technique to be adjusted according to the relative costs of false positives and false negatives (i.e., sensitivity and specificity). Thus our method is better designed for detection of known pathogens.

Detecting the presence of one genome may affect the detection of another genome in the sense that, marginally, the proportion Θ_i of a bacterial strain i in the population, will not be independent of $\Theta_{i'}$. That is, some reads may be shared by two genomes so the presence of one genome may be positively associated with the presence of the other genome. Because our method is based on a single joint posterior across genomic sources we can investigate this dependence. We do this for pairs of genomes at the end of Sec. 2 and find that the dependence is local rather than global, in the sense that even though most genomes are independent there are small groups of related genomes that appear to be dependent due to sequence similarity. This type of assessment is not readily provided by non-Bayes methods.

We demonstrate the behavior of our method on two metagenomic samples from the HMP Data Analysis and Coordination Center (DACC). Our method detects bacterial strains that are likely to be present in the two samples, as determined by their marginal posterior probabilities (>0.95). The HMP characterized these samples using a different alignment strategy and reference database, but without a probabilistic assessment of the reliability of identification. Our conclusions concur broadly with those found by the HMP [18, 19]. We attribute many of the differences in detected strains to differences in alignment methods and reference databases, as well as to the relative costs we assigned implicitly to sensitivity and specificity.

In Section 2 we present a Bayes framework for strain detection and assessment of strain dependence. We present the application of our method to HMP data in Section 3, and compare our results to those provided by the HMP. We also present a measure of dependence for each pair of strains detected. In Section 4 we discuss various aspects of the overall analysis, including evidence for the presence of a genomic source not included in the reference database. Computational details are given in Appendix A.

2. METHOD

We model the sample of N genomic reads $(r_1, r_2, \ldots, r_N)$ as originating from a mixed population of M possible bacterial genomes and an additional genomic 'source' not represented in the reference database (for a total of $M+1$ genomic sources). The reference database is represented as a set of M genome sequences $(g_1, g_2, \ldots, g_M)$. In most bacterial sequence databases, the genome sequence for a bacterial strain may be represented by a collection of sequences each representing a part of the genome, i.e., a chromosome, plasmid, or other DNA scaffold. Our read mappings are performed at the level of each partial genome sequence, and then the results are combined to the strain level for probabilistic analyses. (We do not concatenate the sequences to the strain level prior to analysis because reads mapping across concatenation points may not be biologically plausible). If all the K_i's are non-negative integers, the probability of observing $K_i = k_i$ reads aligning to reference genome/source g_i for $i = 1, \ldots, M+1$ is assumed to follow a multinomial distribution with parameter $\theta = (\theta_1, \theta_2, \ldots, \theta_M)$, i.e.,

$$\begin{aligned} &w(K_1 = k_1, K_2 = k_2, \ldots, K_{M+1} = k_{M+1}|\theta) \\ &= \binom{N}{k_1, k_2, \cdots, k_{M+1}} \theta_1^{k_1}\theta_2^{k_2}\cdots\theta_{M+1}^{k_{M+1}}. \end{aligned}$$

Of course, not all the K_i's are non-negative integers and we correct for this shortly.

The multinomial assumes that the reads are independent when of course they are not: Observing a read from a given source will increase the probability of observing other reads from the same source. Since the exact dependence structure is unknown and might be essentially unknowable in practice, we take this dependence into account as a scaling factor γ_i on the observed read counts for genomic source i, i.e.,

$$k_i^{**} = \gamma_i k_i$$

Again, the k_i^{**}'s are not in general non-negative integers so, as a convenient approximation, we replace them with

$$k_i^* = \mathsf{round}(k_i^{**})$$

where $\mathsf{round}(x)$ means we round x to the nearest integer. The difference in end results from using k_i^{**} or k_i^* are negligible.

If reads from genomic source i are perfectly dependent, i.e., if any one of them occurs it is equivalent to all of them occurring (apart from reads shared with other strains) then these reads provide information proportional to the length of the read only, so we have $\gamma_i = l_r/l_{g_i}$, where l_r is the length of a read and l_{g_i} is the length of genome i. Analogously the case of complete independence corresponds to $\gamma_i = 1$. So, it is reasonable to choose

$$\gamma_i \in [l_r/l_{g_i}, 1],$$

for $i = 1, \ldots, M$. That is, γ_i is chosen to reflect the dependence structure in the data. (For $i = M+1$ we set $l_{g_{M+1}} = \bar{l}_g$ where $\bar{l}_g$ is the mean of the lengths of the genomes in the reference database.) We separately investigate whether the degree of dependence encapsulated by the γ_i's is roughly consistent with the degree of dependence suggested by a separate measure of dependence (see (2) below). In our examples here all reads are of the same length $l_r \equiv 100$ but the above equations generalize easily to other cases.

In the examples to follow, we examine the relationship between the choice of γ_i and the rate of detection, choosing γ_i by putting all of them on a common scale and plotting the number of genomes detected as a function of the common scale. Since the resulting curve is increasing in the scaling on the γ_i's, we choose the scaling value to be the one that identifies the knee in the curve. The curves we get are smooth and steeply rising on a small interval of the form $[0, \epsilon_0)$ but past a certain ϵ_0 they rise more slowly and flatten out. Thus, the knee in the curve (sometimes called the elbow) appears to be well-defined in practice. That is, using the scaling value, γ_i is chosen to balance the costs of over- and under-detection: We want γ_i low to protect against over-detection but high to protect against under-detection. The point at which the curve appears to change character is a transition point from being confident there are few false positives and being confident there are few false negatives.

Using the knee in the curve as a technique to identify an optimal point is a standard technique in some contexts, e.g., in choosing the number of principal components to use in a principal component regression analysis (called a scree plot), the number of clusters to use in a clustering (see [30]), or choosing a classifier (choose the classifier represented by the point on the ROC curve closest to (0,1)). However, it is not common to define the knee formally and its reliability in the sense of estimating something meaningful is a sort of 'folk theorem'. Recently, [17] provided a summary of the debate surrounding the use of the knee in the curve admitting that some regard the knee in the curve as ill-defined or not meaningful. However, [31] had already proposed formalizing the concept by using the curvature function of a curve in the plane and [6] used this definition — equivalent to finding the point of smallest radius of curvature — to estimate proportions of a mixed sample. More recently, [5] simplified this definition to a second derivative condition and verified consistency in a micro-array context. Although their proof does not directly apply to the present NGS setting, it suggests that the knee in the curve, as used here, is a well defined and meaningful concept. Moreover, the results from our examples below are not inexplicably far from related findings. Hence we suggest that, even in the absence of formality, using the knee in the curve is a reasonable way to choose γ_i and the curves we use suggest that there is some meaning to the γ_i's chosen.

In metagenomic contexts where multiple strains of the same species may be present, it is common for some reads to align to more than one genome (*nonunique* reads) due to sequence similarity among genomes. This has been handled in different ways across methods, from discarding reads which map to several genomes to treating the true source as 'missing' and using an expectation-minimization (EM) approach to infer the source genome [11]. A priori we prefer not to discard nonunique reads as they do provide information, albeit limited, but using an EM approach will not scale to thousands of reference genomes, particularly under a non-conjugate prior. We choose instead to treat nonunique reads as providing *fractional* information, i.e., if read r_k aligns to genomic sources g_i and $g_{i'}$ then we allocate the equivalent of 1/2 of a read to k_i and $k_{i'}$. More generally, if r_k aligns to I genomic sources $(g_{i_1}, \ldots, g_{i_I})$ then r_k provides $(1/I)^{th}$ of a read to each of $(k_{i_1}, \ldots, k_{i_I})$.

A separate issue of the reads, apart from non-uniqueness or dependence, is the quality of the reads in the sense of phred scores, see [10]. Roughly, a phred score is an assessment of the reliability of the sequencing on a nucleotide-by-nucleotide basis. Assuming phred scores are good indicators of the reliability of the sequencing, it is an open question whether or not to include all reads. One might argue that low quality reads should be filtered out and only high quality reads used so as to be sure that one will not be misled by low quality data. On the other hand, one might argue that including the low phred score reads will improve the inference analogous to the way a collection of weak learners may combine to provide good inference in boosting or other model averaging techniques. For the two data sets we analyze in this paper, we present some results using all the reads but focus attention on results based on the high quality reads (phred score ≥ 19). An alternative we have not implemented is to weight the fractional reads by their phred scores; we expect this would make little difference given that by filtering out at phred score 19 we are already eliminating well over 50% of the reads.

To complete the specification of how the posterior can be found, we choose the prior distribution for $\Theta = (\Theta_1, \ldots, \Theta_{M+1})$ to be a conjugate Dirichlet distribution with hyperparameter $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_{M+1})$, i.e.,

$$p(\theta \mid \alpha) \propto \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \cdots \theta_{M+1}^{\alpha_{M+1} - 1}$$

and yielding a posterior distribution $W(\theta \mid k^*, \alpha^*)$, where $k^* = (k_1^*, k_2^*, \ldots, k_{M+1}^*)$, which is also Dirichlet with parameters $\alpha^* = (\alpha_1^*, \ldots, \alpha_{M+1}^*) = (\alpha_1 + k_1^*, \alpha_2 + k_2^*, \ldots, \alpha_{M+1} + k_{M+1}^*)$. In this formulation the posterior marginal distribution for each Θ_i is $\mathsf{Beta}(\alpha_i + k_i^*, \sum_{j \neq i}(\alpha_j + k_j^*))$. The hyperparameters $(\alpha_1, \ldots, \alpha_{M+1})$ can be seen as representing 'pseudo counts', or the number of reads we expect to come from each genomic source a priori. Indeed, the parameters in the Beta distribution indicate that α_i must be on the same scale as k_i^*. Two natural choices for the hyperparameters are the following. First, one might set all α_i to be the same constant by invoking the Principle of Insufficient Reason and then choose that constant to be one on the grounds that the smaller the α_i the more influence the data will have. Second, one might choose the α_i's to be a fraction of the size of the g_i's: Bigger g_i's should get higher α_i's on the grounds that given uniform sampling there is a higher chance of reads from larger versus smaller genomes. It turns out that the second method is hard to formulate without making assumptions about either the expected sequencing coverage (or expected genomes present). In most applications this information is not available, so we resort to the first (simpler) method as a reasonable default.

Given the posterior we can do the hypothesis testing. The Bayes test for

$$\begin{aligned} &\mathcal{H}_{0,i} : g_i \text{ is not in the mixed population} \\ (1) \qquad &vs.\ \mathcal{H}_{1,i} : g_i \text{ is in the mixed population,} \end{aligned}$$

is based on

$$W(\theta_i > t \mid k^*, \alpha^*) > 1 - \tau$$

where t and τ are specified thresholds. One natural choice for t is $1/(M+1)$ because it represents the assumption that a priori each genomic source is equally likely. That is, if all genomic sources are equally likely to be present and reads are generated at random, then the proportion of reads from each genomic source should be $1/(M+1)$. Of course this threshold does not take into account the varying lengths of the reference genomes, i.e., k reads from a small genome is more evidence of presence than the same number of reads from a large genome. We can adjust for this by using

$$t_i = l_{g_i} / \sum_{j=1}^{M+1} l_{g_j}$$

as the threshold for genomic source i; each threshold is weighted by the length of the associated genomic source. Moreoever, τ may be found by back solving from the requirement that the Bayes Factor in favor of $\mathcal{H}_{1,i}$ be greater than, say, 3.2 [20]. Thus given the specification of the prior and likelihood, one can form the posterior easily, and perform the M Bayes tests in (1).

Aside from being fairly straightforward to explain and compute, the Bayes framework can also be used to assess the dependence between genomic sources, i.e., the degree to which the appearance of the presence of one genome influences the appearance of the presence of another genome. This can be done for any genome pair by comparing the joint posterior marginal $W(\theta_i, \theta_j \mid k^*, \alpha^*)$ for (θ_i, θ_j) with the product of the univariate posterior marginal distributions $W(\theta_i \mid k^*, \alpha^*)$ and $W(\theta_j \mid k^*, \alpha^*)$ for θ_i and θ_j. In other words,

$$\begin{aligned} f(t_i, t_j \mid k^*, \alpha^*) &= W(\Theta_i > t_i, \Theta_j > t_j \mid k^*, \alpha^*) \\ (2) \qquad &\quad - W(\Theta_i > t_i \mid k^*, \alpha^*) W(\Theta_j > t_j \mid k^*, \alpha^*) \end{aligned}$$

can provide an assessment of dependence. Effectively, $f(t_i, t_j \mid k^*, \alpha^*)$ measures how much the knowledge about the presence of one genome affects the uncertainty about the presence of the other. Because α and k^* are on the same scale, and all of the θ_i's are on the same scale, $f(t_i, t_j \mid k^*, \alpha^*)$ remains a meaningful assessment of dependence even when generalized to three or more genomes.

3. APPLICATION TO HMP DATA

The Human Microbiome Project (HMP) is an NIH-funded research initiative aimed at characterizing the microbial communities found at various sites of the normal human body. The first phase of the HMP (2007–2012) focused on the characterization and composition of the microbial communities which inhabit major mucosal surfaces of the healthy human body. The Project conducted whole metagenome DNA sequencing on biological samples from hundreds of individuals using Illumina technology, and performed metagenomic analyses on these samples, with a series of associated publications in 2012. The metagenomic analyses involved data pre-processing, read assembly and read mapping, and metabolic and functional profiling of samples. The second phase of the HMP (2013–2015) is focused on characterizing the biological properties of the microbiome and host in several disease contexts. The HMP provides metagenomic data and tools for the research community, including NGS sequencing data and analysis pipelines. The HMP resources are available at http://hmpdacc.org and described in associated publications [18, 19]. A complete overview of the HMP data analysis process is at http://www.hmpdacc.org/START/.

We selected two samples from the HMP website for analysis, sample SRS105072 (mid-vaginal) and sample SRS014468 (saliva), which represent relatively low and high diversity bacterial communities, respectively. General descriptions of the collection and processing of the samples to generate the data are described at the HMP data portal noted above. Reads aligning to the human genome have been previously removed by the HMP, and the remaining reads are believed to be largely bacterial or viral in origin. Both samples consist of paired-end 100 base pair reads; sample SRS015072 consists of 495,256 reads while sample SRS014468 consists of 1,159,503 reads. We consider both unfiltered data and data once the reads have been filtered for quality (phred $\geq$ 19); the filtered data consists of 322,541 and 202,487 paired-end reads, respectively.

Next we must select, obtain, and preprocess a reference database of bacterial genomes. Our reference set consists of all the bacterial genomes from the Integrated Microbial Genomes (IMG, version 4.0) database [24]. We prepare all of the genomic files for alignment by indexing the files using the indexing software (bowtie2-build) of the Bowtie2 [21] aligner. Further details about the aligner and indexing are provided in Appendix A.

Given the sample reads and the reference database we align the reads to the reference database and adjust the number of reads aligning to each genome for non-unique reads (fractional read counts). This involves counting, for each read, the number of genomic files to which the read aligns, and adjusting the read counts for each genomic file accordingly (i.e., if a read aligns to n genomic files, the read contributes $1/n$ reads to the total read count of each genome file). In this way we generate k_i for each g_i in the reference database.

3.1 Choice of dependence factor γ

As described in Section 2 we used read counts k_i^*, the read counts k_i adjusted for the dependence among reads from the same genomic source and rounded to the nearest integer. As the nature of the dependence is unknown, we represented this dependence by a factor γ_i which was specific to each genome and could be estimated from the data. We examine the graph of the number of genomic sources detected in the data as a function of the scaling of the γ_i's, where each unscaled $\gamma_i \in [l_r/l_{g_i}, 1]$. Such graphs are qualitatively similar to the graphs shown in Figure 1 for the unfiltered and filtered data. In these graphs, for a given x-axis value h, we plot the number of genomic files detected when the h^{th} largest threshold for each file is used, i.e., we use γ_{ih} for each genome i where the thresholds $(\gamma_{i1}, \ldots, \gamma_{i1000})$ for genome i form a uniform partition of $[l_r/l_{g_i}, 1]$ of size 1,000. Note that the range for γ_i (and therefore the spacing of the thresholds) depends on i, i.e., the grid is not uniform across i, even though the number of thresholds is constant over i.

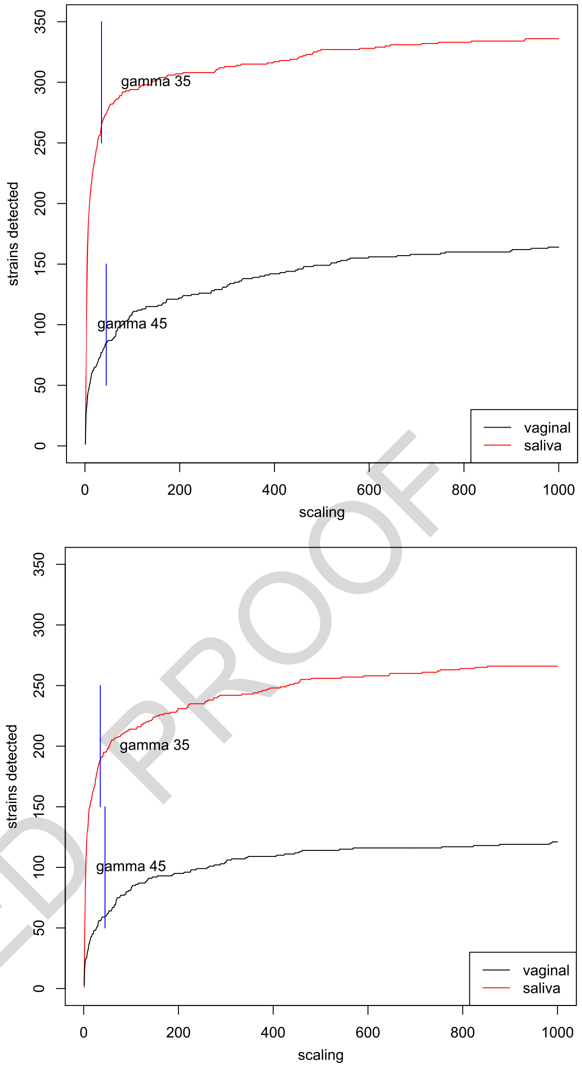

Figure 1. Number of genomic files detected as a function of γ; unfiltered (top) and filtered (bottom) data.

An example may clarify this. Consider three strains g_1, g_2 and g_3 and suppose $\gamma_1 \in [100/1000, 1]$, $\gamma_2 \in [100/10^6, 1]$, $\gamma_3 \in [100/500{,}000, 1]$ where the read length $l_r = 100$ in all three cases and the genome lengths are 1,000, 10^6, and 500k, respectively. Each of the three intervals is partitioned into 1,000 subintervals with endpoints, say $g_{1,1}, \ldots, g_{1,1000}$, $g_{2,1}, \ldots, g_{2,1000}$, and $g_{3,1}, \ldots, g_{3,1000}$. These do not coincide from genome to genome, but their labels, i.e., the indices of their order, do. That is, we can associate, say, the v-th intervals for each genome, i.e., form 1,000 triples $(g_{1,v}, g_{2,v}, g_{3,v})$, so that even though the subintervals are different from genome to genome it is only the ordering that matters. In this example, there are 1,000 triples and each triple corresponds to the possible value of a vector of the form $(\gamma_1, \gamma_2, \gamma_3)$. It is vectors like these that are used in Fig. 1, except the length is 5,168 or the number of reference strains.

In the graphs of Fig. 1, we can see a common pattern: Rapid increase followed by a leveling out. This reflects an initial rapid increase in the number of genomes detected, followed by the more gradual inclusion of further genomes as γ increases. In order to balance false-positive and false-negative findings, it seems reasonable to select γ to represent the point of this qualitative change, i.e., the change from rapid inclusion to slow inclusion. Often this is called finding the knee in the curve; it is a standard procedure in principal components analysis and receiver operating characteristic curves in classification, among other settings. For the samples here this leads us to dependence factors of γ_{i45} (the 45^{th} largest of 1,000 factors for each i) for the mid-vaginal sample and γ_{i35} (the 35^{th} largest of 1,000 factors for each i) for the saliva sample. These values are the same for the filtered and unflitered data.

3.2 Results of bacterial strain detection

Given a fixed scale value for each sample we can do the Bayes testing as presented in (1). That is, we infer that genomic source i is present if

$$W(\Theta_i > l_{g_i} / \sum_{j=1}^{M+1} l_{g_j} \mid k^*, \alpha^*) > 1 - \tau.$$

Note that these same posterior probabilities were used to generate the graphs in Figure 1 and select the γ_i's. That is, we are using the data twice — first to estimate the nuisance parameter γ and then to find the actual posterior. This double usage of the data is necessary because within the Bayes paradigm one cannot evaluate bias. Using the data twice is one way to compensate when the estimate of the nuisance parameter can be regarded as helping to ensure the model is fit to the data well. In our examples here, the sample size per parameter is large enough that this is unlikely to be a problem: For the mid-vaginal sample there were 405k reads, corrected to .45 × 405k = 182k independent reads for about 5k parameters giving about 182k/5k, or 36 data points per parameter. Overall, this reinforces our interpretation of the scaling as representing a trade-off between false positives and false negatives.

In a further pragmatic correction, we did not test any bacterial strain with less than five reads aligning. This is a simple way to ensure that the results would not be prior-driven. Since the α_i's were all one and represented 'virtual reads' using a cutoff of five reads seemed reasonable. (Using a cutoff of 10 reads meant that we lost some strains that were closely related to other strains detected and this seemed counter-productive.)

For the mid-vaginal sample (unfiltered) we detect 85 bacterial strains representing 47 bacterial species. The mean (median) read count per strain was 4,554 (735) reads. This reflects a highly skewed distribution with a fairly wide range; this is partly explained by the scaling of the thresholds t_i for the size of the genomes. The HMP reported 29 strains as present, representing 15 bacterial species. Of these, we detected 27/29 strains and 13/15 species. The two species/strains that we fail to detect, Sphingopyxis alaskensis RB2256 and Stenotrophomonas maltophilia K279a, are reported by HMP to have relatively low sequence coverage (depth/breadth of 0.020/1.99 and 0.010/1.21, respectively). For the filtered case we detect 63 bacterial strains representing 40 bacterial species. The mean (median) read count per strain was 5,730 (633) reads. Comparing with the HMP findings we detected 24/29 strains and 13/15 species, so slightly lower overlap relative to the unfiltered data. Many of the strains we detected that were not reported by the HMP belong to species detected by the HMP and the consensus is even stronger at the genus level; this may partly reflect differences in the alignment method and reference databases. The overlap between the lists of detected strains/species based on the unfiltered and filtered data consists of 61 strains and 38 species; see Figure 2. The discrepancy between the results based on the unfiltered and filtered data is not surprising if we consider that filtering removed 59.1% of the reads from the mid-vaginal sample.

For the saliva sample (unfiltered) we detect 139 bacterial strains representing 94 bacterial species. The mean (median) read count per strain was 348 (21) reads; this is much lower than for the mid-vaginal sample due to the increased complexity of the population. The HMP reported 140 strains as present, representing 105 bacterial species. Of these, we detected 50/140 strains and 46/105 species. For the filtered case we detect 91 bacterial strains representing 75 bacterial species. The mean (median) read count per strain was 383 (26) reads. Comparing with the HMP findings we detected 41/140 strains and 41/105 species. As for the mid-vaginal sample, many of the strains we detected that were not reported by the HMP belong to species detected by the HMP, and the consensus is even stronger at the genus level. However, our results and the findings of the HMP are more disparate due to the increased complexity of the population from which the sample was taken. The overlap between the lists of detected strains/species based on unfiltered and filtered (ignoring the overlap with HMP) data consists of 47 strains and 36 species, if the HMP strains/species are included this increases to 83 strains and 69 species; see Figure 2. As noted above, the discrepancy between the results based on the unfiltered and filtered data is expected as filtering removed 72.2% of the reads.

To provide an alternative perspective on our findings we plotted the strains detected in the filtered data as a function of genome size and sequencing depth; see Figure 3. The overwhelming majority of strains have small genomes and low sequencing depth, which contributes to the overall uncertainty

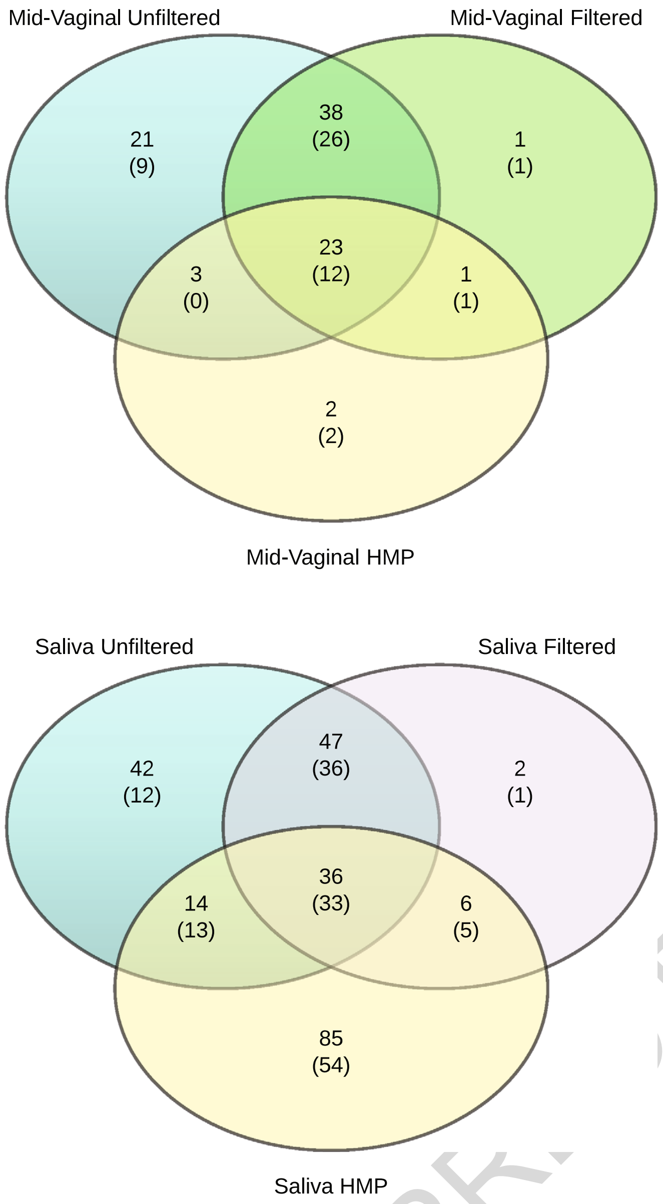

Figure 2. Summary of the overlap among the filtered, unfiltered, and HMP results. Top: mid-vaginal. Bottom: saliva. The numbers without parentheses are the number strains detected when a minimum of 5 reads is required; the smaller numbers without parentheses are the number species detected again when a minimum of 5 reads is required.

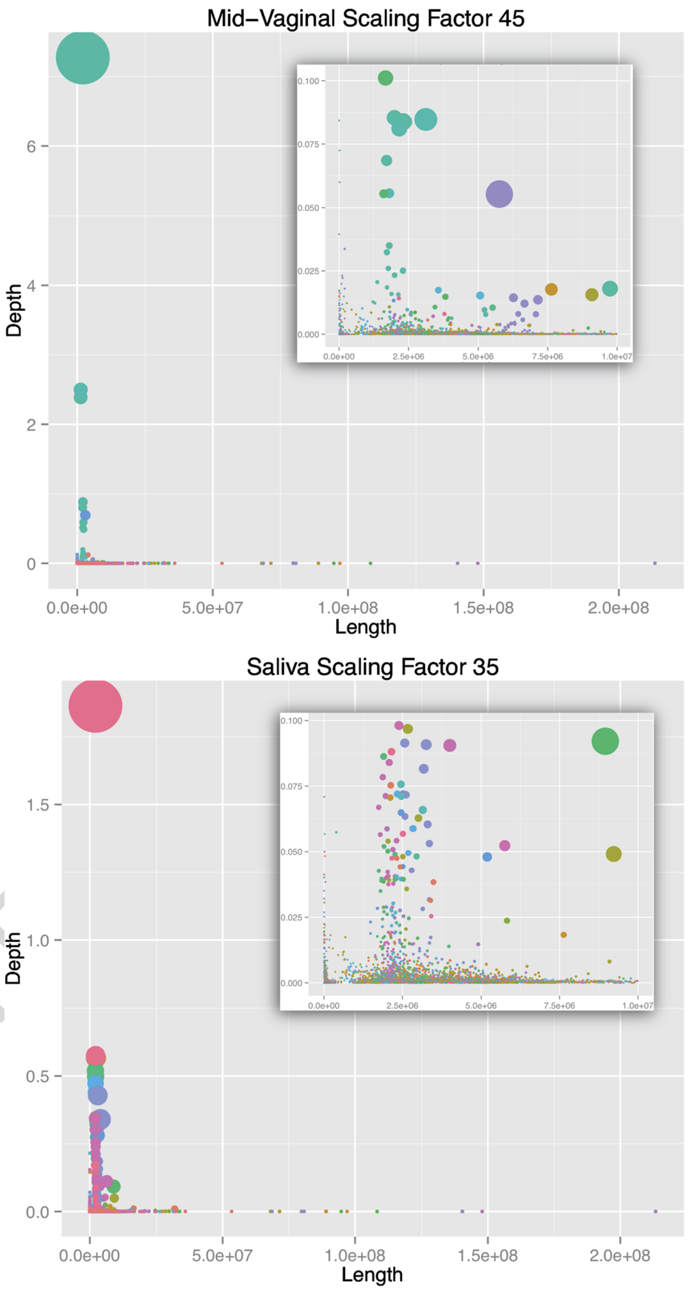

Figure 3. Strains detected in filtered data. Top: mid-vaginal at γ_{i45}. Bottom: saliva at γ_{i35}. The x-axis is the genome length in base pairs and the y-axis is the sequencing depth. The dot size is proportional to the posterior marginal detection probability and the dot color represents bacterial genus. The inset highlights the lower left corner of the graph.

in population composition. Note the difference in scales for sequencing depth; because fewer strains are present in the mid-vaginal sample is it possible to achieve higher sequencing depth. This also reflects how, after filtering, the saliva sample consisted of fewer reads than the mid-vaginal sample. The insets of the two graphs highlight the phylogenetic diversity of the saliva sample (many genera represented) relative to the phylogenetic depth of the mid-vaginal sample (many strains of specific genera represented).

We comment that in our analysis of the whole metagenome DNA sequencing data from HMP we have assumed a priori that each genome in a biological sample has the same probability of being sequenced, and more abundant genomes have a higher probability of being sequenced than less abundant genomes. This is a standard assumption even though it is at best only approximately true. Under it, more abundant genomes will generate relatively more sequencing reads, which our method will map to the relevant reference genomes. Thus we expect that, given two genomes of similar sizes, our method will assign higher posterior probabilities of presence to the genome with higher abundance. This is not to say that abundance and posterior probability are directly correlated, as genome size and depth of sequencing also play a role in determining the probabilities of detection (see Figure 3). However, we do expect those bacteria with larger genomes and higher abundance in any given sample to yield higher posterior probabilities of presence. Otherwise put, there is a threshold of abundance and sequencing coverage that must be satisfied in order for our method to detect any specific genome. This is a property of detection methods in general.

Note that there is a tradeoff between coverage and abundance in terms of detection. If a genome with low abundance has high enough coverage or a genome has mow coverage but high enough abundance, it will be detected. The optimal case for detection is high abundance and high coverage. It is only when both abundance and coverage are too low that a genome that is present will fail to be detected.

Our model allows for the detection of a genomic source not represented in the reference database. This source is represented by our $M + 1$st genomic category. In both the mid-vaginal and the saliva datasets this category was detected with posterior probability > 0.99. Since human reads were pre-screened from both datasets we conclude that a genomic source not in the reference database, of non-human origin, is present in both datasets. In order to identify this source we could align the reads associated with this category to other genomic databases, such as those for viruses and other eukaryotes, and use the method presented here to determine presence/absence of specific sources. An alternative with the unfiltered data is that the reads detected to be in category $M + 1$ may merely be such low quality reads that they do not match to anything in our database. In many cases it is not a priori clear which case — low quality reads or missing reference genomes — are represented by category $M + 1$.

3.3 Pairwise dependence between strains

It was argued in Sec. 2 that using an appropriate scaling factor γ could be used to correct for any dependence in reads, hence making it reasonable to use a multinomial likelihood. To verify that this is the case, we generated histograms of the joint probability of detection minus the product of probabilities of detection for all the genomes detected. That is, we plotted the values

$$\begin{aligned} &W(\Theta_i > t_i, \Theta_j > t_j \mid k^*, \alpha^*) \\ &\quad - W(\Theta_i > t_i \mid k^*, \alpha^*) W(\Theta_j > t_j \mid k^*, \alpha^*) \end{aligned} \tag{3}$$

where i, j ranged from 1 to the number of genomes detected in each case (filtered, unfiltered; cutoff of five reads minimum and $W(\Theta_i > t_i \mid k^*, \alpha^*) > 1 - \tau$; and mid-vaginal, saliva). Expression (3) is a measure of dependence because it is zero when the i-th and j-th genomes are independent and as it increases in absolute value it indicates higher dependence; expression (3) is essentially the strong mixing condition (sometimes called α-mixing).

As a representative example, Fig. 4 shows the histograms from calculating (3) for the filtered data. The upper panel shows the results for the mid-vaginal data and the lower panel shows the results for the saliva data. It is seen that for the mid-vaginal data there is a large spike at zero. Indeed, a large majority of the differences in (3) are smaller than 0.1 in absolute value; the tail on the right merely indicates the association is generally positive. This means that the joint probability is higher than the product of the marginal probabilities so that detecting one genome makes detecting some other genomes more likely. For the saliva data, it is seen that the concentration around zero is slightly stronger than for the mid-vaginal data, and the tail is again to the right, suggesting a positive association between genomes. Note that the direction of dependence is the same for both cases, intuitively reasonable since detecting one genome increases the probability of detecting similar genomes. An interesting difference is due to the complexity of the data set. Between the mid-vaginal and saliva data sets the vertical scales differ by a factor of ten, because the saliva data set is so much more diverse, i.e., the number of pairs of strains increases with more strains present. In addition the saliva data set contains stronger pairwise dependencies, as seen by the range of the x-axes for the two plots.

A separate question from how much association seems to be present is to ask what form it takes: Which genomes seem to be dependent on which other genomes? We address this question by using network dependence plots [37]; see Fig. 5 for the same cases as in Fig. 4, i.e., filtered data. We used a cutoff of 0.03 for the mid-vaginal data set and 0.06 for the saliva data set so as to make the size of the network dependence graphs roughly equal. (However, the number of pairs with strictly positive dependencies in the saliva data set is 12,816, much larger than 1,364 for the mid-vaginal data set; see the y-axis scales on the respective histograms in Fig. 4.) Now, the numbers of links in the two networks are not too different — 74 for mid-vaginal and 62 for saliva. Relative to the network for the mid-vaginal sample, which shows 58 individual strains and 19 genera, the saliva sample shows fewer individual strains (11) and fewer genera (15). This makes it appear that the mid-vaginal data set is more phylogenetically diverse than the saliva data set, however, this is an artifact of the cutoffs: If the cutoff .06 were used for the mid-vaginal data set, its network would have no links. Note also that the unspecified $M + 1$ category appears in the mid-vaginal dependence network (aqua colored). Overall, this shows that there are dependencies, however slight, whose structure may be of interest.

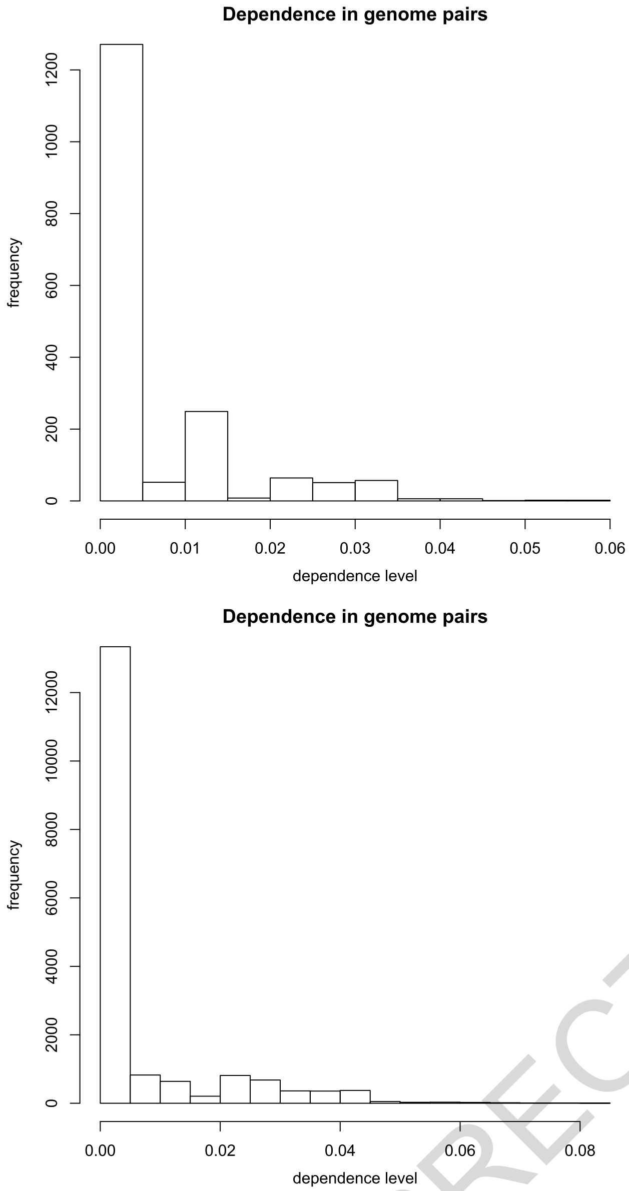

Figure 4. Diagrams of the pairwise dependence between the parameters in the posterior distribution for the strains detected (filered cases). Top: mid-vaginal. Bottom: saliva.

4. DISCUSSION

We have presented a Bayesian approach to statistical strain detection from bacterial metagenomic samples generated by next generation sequencing. Such samples have been generated by multiple research projects including the Human Microbiome Project [18] and the TerraGenome consortium [38]. Our method uses posterior marginal probabilities to detect specific bacterial strains, and quantifies the dependence between pairs of strains by comparing the joint probability of detection to the product of the marginal probabilities of detection. The threshold for detecting the presence of a genome is chosen to be proportional to the length of the genome, providing an automatic adjustment for genomic length. In order to incorporate the dependence among reads from the same genome, we allow for a scaling factor on the read counts for each genomic source; this scaling factor is a nuisance parameter whose estimate takes into account both the read length and the length of the reference genome. The Bayes paradigm is also able to quantify the evidence in favor of the presence of an unknown genomic source, i.e., a source of genomic material that is not present in the reference database.

The scaling factor on the read counts can be selected to provide a balance between false detection and failure to detect, i.e., false positives and false negatives. This is an advantage over existing approaches such as Pathoscope [11] which, as the authors note, has a tendency toward parsimony and can miss one of more similar substrains. In the presence of ambivalent information, i.e., reads which align to more than one genomic source, we share this information across the relevant sources and quantify it probabilistically. In our opinion this provides more information than discarding 'non-unique' reads or only providing the 'best choice' for mapping non-unique reads.

Note that the number of strains present for which Pathoscope was demonstrated effective ranged from three to 30; the number of strains in their reference set was 131. However, it will be very difficult to scale Pathoscope up to larger numbers of strains present or in the reference set because Pathoscope is based on the EM-algorithm for which both running time and convergence diagnostics will be problematic in general. By contrast, in our examples we had an unknown number of strains present, and over 5,000 strains in the reference set. Moreover, it is clear that our procedure will scale up readily to even higher numbers of strains present or in the reference set — irrespective of how similar or dis-similar the strains are.

Note that our estimate of the scaling factor is not Bayes, so the overall procedure is empirical Bayes. While philosophically impure and a limitation of the method, it is probably not a problem in practice — at least when the sample size is large enough. Here, we have linked the parameters γ_i into a single parameter γ used to adjust for dependence. For the smaller sample we have 405k reads and about 5k strains and our correction for the dependence was a factor of .45. Thus it's as if we had $.45 \times 405\text{k}/5\text{k} = 36$ independent data points (reads) per parameter. A similar calculation can be done for the larger sample. Aside from pathological cases this is usually more than enough for posteriors to exhibit convergence. Hence, one expects good posterior behavior since the sources of variability (e.g., the dependence) that have been included in the modeling are typically going to have a much greater effect than those that have been neglected (e.g., estimating γ).

It is important to note that our technique does not minimize false positives or false negatives; it chooses a balance between these two extremes. If there is ambivalent information then our use of fractional reads means it is shared and

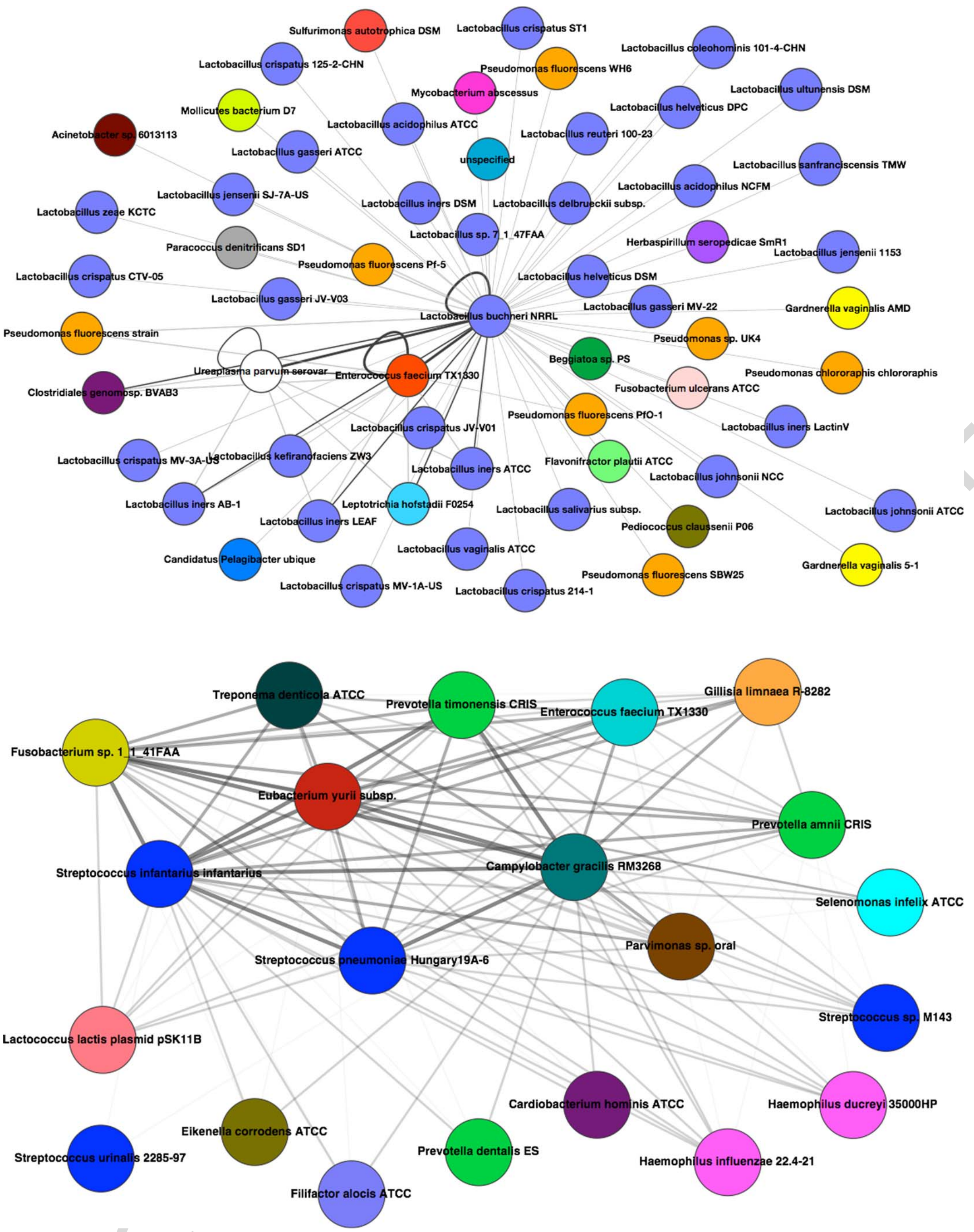
Sulfurimonas autotrophica DSM
Lactobacillus crispatus ST1
Lactobacillus coleohominis 101-4-CHN
Lactobacillus crispatus 125-2-CHN
Pseudomonas fluorescens WH6
Mycobacterium abscessus
Lactobacillus ultunensis DSM
Mollicutes bacterium D7
Lactobacillus helveticus DPC
Acinetobacter sp. 6013113
Lactobacillus acidophilus ATCC
Lactobacillus reuteri 100-23
Lactobacillus gasseri ATCC
unspecified
Lactobacillus sanfranciscensis TMW
Lactobacillus jensenii SJ-7A-US
Lactobacillus acidophilus NCFM
Lactobacillus iners DSM
Lactobacillus delbrueckii subsp.
Lactobacillus zeae KCTC
Lactobacillus sp. 7_1_47FAA
Paracoccus denitrificans SD1
Herbaspirillum seropedicae SmR1
Lactobacillus jensenii 1153
Pseudomonas fluorescens Pf-5
Lactobacillus crispatus CTV-05
Lactobacillus helveticus DSM
Lactobacillus gasseri JV-V03
Lactobacillus gasseri MV-22
Gardnerella vaginalis AMD
Lactobacillus buchneri NRRL
Pseudomonas fluorescens strain
Pseudomonas sp. UK4
Beggiatoa sp. PS
Ureaplasma parvum serovar
Enterococcus faecium TX1330
Pseudomonas chlororaphis chlororaphis
Clostridiales genomosp. BVAB3
Fusobacterium ulcerans ATCC
Pseudomonas fluorescens PfO-1
Lactobacillus crispatus JV-V01
Lactobacillus iners LactinV
Lactobacillus crispatus MV-3A-US
Lactobacillus kefiranofaciens ZW3
Flavonifractor plautii ATCC
Lactobacillus iners ATCC
Lactobacillus johnsonii NCC
Lactobacillus iners AB-1
Leptotrichia hofstadii F0254
Lactobacillus salivarius subsp.
Lactobacillus johnsonii ATCC
Lactobacillus iners LEAF
Pediococcus claussenii P06
Lactobacillus vaginalis ATCC
Candidatus Pelagibacter ubique
Gardnerella vaginalis 5-1
Pseudomonas fluorescens SBW25
Lactobacillus crispatus MV-1A-US
Lactobacillus crispatus 214-1
Treponema denticola ATCC
Gillisia limnaea R-8282
Prevotella timonensis CRIS
Enterococcus faecium TX1330
Fusobacterium sp. 1_1_41FAA
Eubacterium yurii subsp.
Prevotella amnii CRIS
Streptococcus infantarius infantarius
Campylobacter gracilis RM3268
Selenomonas infelix ATCC
Parvimonas sp. oral
Streptococcus pneumoniae Hungary19A-6
Streptococcus sp. M143
Lactococcus lactis plasmid pSK11B
Cardiobacterium hominis ATCC
Haemophilus ducreyi 35000HP
Eikenella corrodens ATCC
Streptococcus urinalis 2285-97
Prevotella dentalis ES
Haemophilus influenzae 22.4-21
Filifactor alocis ATCC

Figure 5. The strains detected with the strength of the dependence indicated by the lines connecting the vertices. Vertices representing different strains from the same genus are given the same color. Top: mid-vaginal. Bottom: saliva.

quantified probabilistically rather than optimized to give a 'best guess' or 'best choice'. As a separate point, while our method exploits conjugacy, it is obvious how to extend our method to the non-conjugate prior case: It is enough to be able to obtain the univariate posteriors, one for each genome, in order to do the Bayes tests.

A possibly controversial feature of our method is that it does not use a multiple comparisons correction. In fact, to be very strict about it, the Bayes multiple testing problem has not really been clearly defined. Nevertheless, most authors agree that Bayes procedures have a built in push towards sparsity, sometimes called the 'Ockham's razor' effect, that often obviates the need for explicit multiple comparisons corrections. Hence, many authors agree that there are many cases where Bayes procedures have a multiple correction built in, see [3], [33], [2], [28] and [13], [34]. Some of these authors try to specify conditions where a multiple comparisons correction is necessary, but the conditions they find do not seem to coincide partially because they are studying different settings. Nevertheless, taken together, the line of thought these papers represent, seems to suggest that putting an extra layer of variability on the hypotheses i.e., converting them to a ‘model selection problem’, is a good methodology, even if it is unclear how the structure of the data (independent from test to test or not for instance) or the dependence among the parameters in the joint prior affects this. Indeed, it is not clear when this procedure differs from merely having a prior probability on each null hypothesis being true; the explanation may be that in a real model selection problem the prior has two layers, the within-model prior and the across-models prior, whereas in other testing problems (such as here) there is no analog of the across-model prior to use in marginal tests.

A different approach is taken in [16]. They develop statistics that look essentially frequentist but are asymptotically Bayes in a decision theoretic sense. Indeed, these authors state: ‘thresholding the marginal posterior probability amounts to controlling the positive FDR’ (at least in their setting). Their setting, like some of the others, assumes the data for each hypothesis are independent, the parameters in the tests are different, and there are no nuisance parameters. While somewhat ad hoc, the reliance on decision theory makes sense because Bayes testing is based on the fact that the Bayes factor (or more precisely thresholding the posterior probability) is the Bayes action under generalized zero-one loss. It should be noted that other authors such as [28] also take a decision theoretic approach. On the other hand, this procedure seems difficult to implement, suffers from incoherency (see below) in finite samples, and its asymptotics may make it equivalent to an empirical Bayes procedure such as those criticized in [34].

A different approach again is taken in [34]. They argue that prior selection should be used to effect a multiple comparisons correction in a linear model selection problem so that in essence the built-in multiple correction effect from the Bayes formulation can be exploited. They also criticize empirical Bayes approaches such as used here, in [16], and originating in [14]. However, the framework in [34] really is a model selection problem so unlike some other cases, e.g., [33], the prior on the hypotheses is the across model prior and hence is an essential component of the Bayesian formulation rather than an added construct to combine the hypotheses into one big measure space (the Bayes containment principle). That is, the multiple testing problem does not have to be converted into a model selection problem because it already is one.

At root, there are two ways to justify Bayes testing at least in the simplest cases. One is the well-known decision theoretic criterion posterior risk under the generalized zero-one loss function. The other is via coherency arguments such as originate in [8] and were developed in [12]. The decision-theoretic approach is constructive in that it leads to the posterior probabilty of a hypothesis as the right thing to use even if the threshold depends on the loss function. The optimality of the use of the posterior odds under coherency amounts to saying that any other way of posting odds leads to a certain loss of money by the bookie.

The stance (gingerly) taken here is the following and is supported by the fact that the results are more-or-less in the range one would expect by comparison with the HMP results for strain/species detection. First, each individual univariate hypothesis test should be coherent in the sense of [12] so that means one must use the marginal posterior odds from the single posterior conditioned on all the data. Prior selection to avoid multiple comparisons is relatively undeveloped and from a Bayes persepctive can be good only when one has no other auxilliary information to be built into the prior. However, here, we thought we should invoke a Principle of Insufficient Reason to insist all the α_j's be the same and then set them to one to maximize the effect of the data. Therefore, the only parameter left in the analysis to use in a multiple comparisons correction is the threshold of the posterior probability that in principle comes from the generalized zero-one loss function. In effect, this means taking different loss functions for the different tests. The problem is that pre-experimentally we do not know how to formulate the right generalized zero-one loss function and hence cannot identify the ‘right’ cutoff value for the posterior for each hypothesis. Hence we de facto assumed that all the loss functions were the same and so would lead to the same threshold. Therefore, we merely looked for the largest marginal posterior probabilities using uncorrected thresholds (backformed from requiring the Bayes factor to be greater than 3.2). So, two obvious ways to improve the present analysis would be to bring more subject matter knowledge to bear on the selection of the parameters in the generalized zero-one loss function and the verification that the empirical Bayes method used here really is an approximately fully Bayes method (or has some other feature that makes it reasonable).

ACKNOWLEDGEMENTS

We would like to thank Anthony Fodor, Ph.D., for guidance through the labyrinth of microbiology and metagenomics. The authors were supported by NSF-DTRA grant #NR66853W (DMS-1120404).

APPENDIX A. COMPUTATIONAL DETAILS

Bacterial reference sequences

456,865 whole genome bacterial reference sequences, in FASTA format, were downloaded from the Integrated Microbial Genomes (IMG) database (version 4.0) [24]. The 456,865 reference sequences accounted for 5,168 bacterial references — these included sequences from bacterial genomes and bacterial plasmids. The 5,168 bacterial references were isolated by relying on bacterial taxon names and sequence identifiers obtained from the Genome Browser at the IMG website (http://img.jgi.doe.gov).

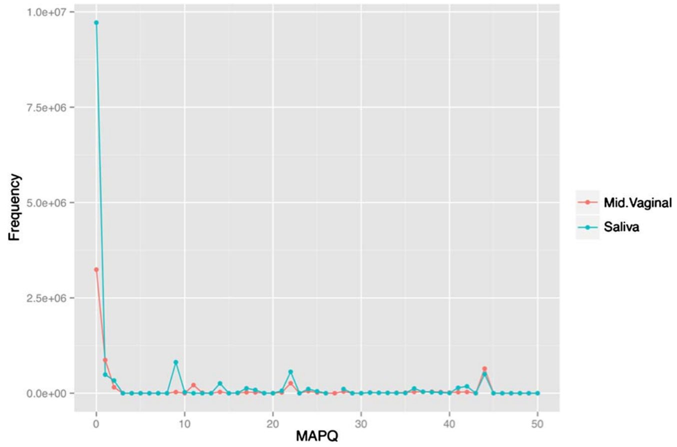

Figure 6. Mapping quality values (MAPQ) for both HMP samples as reported by Bowtie2 where the mapping quality is assessed using phred scores. As noted, phred score ≥19 was used to filter the alignments for downstream analysis.

Metagenomic samples

Two human metagenomic samples were downloaded from the Human Microbiome Project data portal [15]: a Saliva sample (accession SRS014468) and a Mid-Vaginal Sample (accession SRS015072). Both samples are available at http://www.hmpdacc.org/HMSCP/ and consist of 100-bp paired-end reads; The Saliva sample contains 1,159,503 reads, while the Mid-Vaginal sample contains 495,256.

Data processing and local alignment

Both Saliva and Mid-Vaginal samples were aligned to the 456,865 bacterial sequences using the Bowtie2 [21] aligner in the local-alignment mode (reads were not aligned using the traditional end-to-end alignment). The following Bowtie2 command was used:

```
bowtie2 --local -D 20 -R 3 -N 0 -L 20 -i S,1,0.50 --time -f -x -S
```

Previous versions of Bowtie employed a global alignment policy to align reads to a reference. This policy allowed only a certain number of mismatches in the read, and reads were aligned “end-to-end”. Bowtie2’s support of local alignment expands the alignment policy to support the alignment of small chunks in the reads, and allows reads to be aligned without a strict end-to-end policy. The resulting alignments for both samples (Saliva & Mid-Vaginal) were filtered by mapping qualities Samtools (0.1.18) [22]. A phred score of 19 or greater was used as the threshold to filter the alignments by Figure 6.

After filtering, the Saliva sample contained 322,541 paired-reads while the Mid-Vaginal sample contained 202,487 reads.

Post processing

The filtered reads are then analyzed using a custom PERL script that counts the number of hits that a given bacterial reference sequence (genome or plasmid) has. Read hits to a reference are normalized by the number of references that a given hit maps to. Reports at the Strain, Species, and Genus level are then generated.

Received 14 November 2013

REFERENCES

[1] Auer, P. and Doerge, R. (2010) Statistical design and analysis of RNA sequencing data. *Genetics*, **185**, 405–416.

[2] Bayarri, S. and Berger, J. (2005) Multiple testing: The problem and some BAyes and frequentist solutions. http://sisla06.samsi.info/ndhs/ad/Presentations/Overwolfach05.pdf.

[3] Berry, D. and Hochberg, Y. (1999) Bayes perspectives on multiple comparisons. *J. Stat. Planning and Inference*, **82**, 215–227. MR1736444

[4] Brady, A. and Salzberg, S. (2011) PhymmBL expanded: confidence scores, custom databases, parallelization and more. *Nature Methods*, **8**, 367.

[5] Clarke, B. and Clarke, J. (2013) Deconvolution of gene expression from mixed samples. *Submitted*.

[6] Clarke, J., Clarke, B., and Seo, P. (2010) Statistical expression deconvolution from mixed tissue samples. *Bioinformatics*, **26**, 1043–9.

[7] Datta, S., Datta, S., Kim, S., Chakraborty, S., and Gill, R. (2010) Statistical analyses of next generation sequence data: A partial overview. *J. Proteomics Bioinform.*, **3**, 183–190.

[8] de Finetti, B. (1937) La prevision: ses lois logiques, ses sources subjectives. *Ann., L’Inst. Henri Poicare*, **7**, 1–68. MR1508036

[9] Devaraj, S., Hemarajata, P., and Versalovic, J. (2013) The human gut microbiome and body metabolism: implications for obesity and diabetes. *Clinical Chem.*, **59**, 617–628.

[10] Ewing, B., Hillier, L., Wendl, M., and Green, P. (1998) Base calling of automated sequencer traces using phred. Part I: Accuracy assessment. *Gen. Res.*, **8**, 175–185.

[11] Francis, O., Bendall, M., Manimaran, S., Hong, C., Clement, N., Castro-Nallar, E., Snell, Q., Schaalje, G., Clement, M., Crandall, K., and Johnson, W. (2013) Pathoscope: Species identification with strain attribution with unassembled sequencing data. *Genome Res*, **23**, 1721–1729.

[12] Freedman, D. and Purves, R. (1969) Bayes method for bookies. *Ann. Math. Stat.*, **40**, 1177–1186. MR0240914

[13] Gelman, A., Hill, J., and Masanao, Y. (2008) Why we (usually) do not have to worry about multiple comparisons. http://www.stat.columbia.edu/˜gelman/research/unpublished/multiple2.pdf.

[14] George, E. and Foster, D. (2000) Calibration and empirical Bayes variable selection. Biometrika, **87**, 731–747. MR1813972

[15] Gevers, D., Knight, R., Petrosino, J., Huang, K., and McGuire, A. et al. (2012) The Human Microbiome Project: A community resource for the healthy human microbiome. *PLoS Biology*, **10**, e1001377.

[16] Guindani, M., Muller, P., and Zhang, S. (2009) A Bayes discovery procedure. *J. R. Stat. Soc. Ser. B*, **71**, 905–925. MR2750250

[17] Horta, F., Caldera, R., Cortes, R., Carballido, J., and Alpuche, E. (2011) Mathematical model for the optimal utilization percentile in M/M/1 systems. http://arxiv.org/ftp/arxiv/papers/1106/1106.2380.pdf.

[18] The Human Microbiome Project Consortium (2012) Structure, function and diversity of the healthy human microbiome. *Nature*, **486**, 207–214.

[19] The Human Microbiome Project Consortium (2012) A framework for human microbiome research. *Nature*, **486**, 215–221.

[20] Kass, R. and Raftery, A. (1995) Bayes factors. *JASA*, **90**, 773–795.

[21] Langmead, B. and Salzberg, S. (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**, 357–359.

[22] Li, H., Handsaker, B., Wysoker, A., Fennell, T., and Ruan, J. et al. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

[23] Magi, A., Benelli, M., Gozzini, A., Girolami, F., Torricelli, F., and Brandi, M. (2010) Bioinformatics for NGS data. *Genes*, *1*, 294–307.

[24] Markowitz, V., Chen, I., Palaniappan, K., Chu, K., Szeta, E., Grechkin, Y., Ratner, A., Jacob, B., Huang, J., Williams, P., Hunteman, M., Anderson, I., Mavromatis, K., Ivanova, N., and Kyrprides, N. (2012) IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids*, **40**, 115–122.

[25] Mende, D., Sunagawa, S., Zeller, G., and Bork, P. (2013) Accurate and universal delineation of prokaryotic species. *Nature Methods*, *10*, 881–884.

[26] Metzker, M. (2010) Sequencing technologies – the next generation. *Nature Reviews Genetics*, **11**, 31–46.

[27] Neilsen, R., Paul, J., Albrechtson, A., and Song, Y. (2011) Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, **12**, 443–451.

[28] Mueller, P., Parmigiani, G., and Rice, K. (2006) FDR and Bayes multiple comparison rules. *Johns-HopkinsUniversity, Department of Biostatistics Working Paper # 115.*

[29] Salipante, S., Sengupta, D., and Rosenthal, C., et al. (2013) Rapid 16S rRNA Next-Generation Sequencing of polymicrobial clinical samples for diagnosis of complex bacterial infections. *PLoS ONE*, **8**, e65226.

[30] Salvador, S. and Chan, P. (2004) Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. 16-th IEEE International Conference on Tools with Artificial Intelligence, 576–584. See: http://cs.fit.edu/~pkc/papers/ictai04salvador.pdf.

[31] Satopaa, V., Albrecht, J., Irwin, D., and Raghavan, B. (2011) Finding a ‘kneedle’ in a haystack: Detecting knee points in system behavior. 166–171. 31-st International Conference on Distributed Computing Systems. See: http://www1.icsi.berkeley.edu/~barath/papers/kneedle-simplex11.pdf.

[32] Scallan, E., Hoekstra, R., Angulo, F., Tauxe, R., Widdowson, M., Roy, S., Jones, J., and Griffin, P. (2011) Foodborne illness acquired in the United States – major pathogens. *Emerg. Infect. Dis.*, **1**, 7–15.

[33] Scott, J. and Berger, J. (2006) An exploration of aspects of Bayes multiple testing. *J. Stat. Planning and Inference*, **136**, 2144–2162. MR2235051

[34] Scott, J. and Berger, J. (2010) Bayes and empirical Bayes multiplicity adjustment in the variable selection problem. *Ann. Statist.*, **38**, 2587–2619. MR2722450

[35] Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, **9**, 811–814.

[36] Segata, N., Bornigen, D., Morgan, X., and Huttenhower, C. (2012) PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nature Communications*, **4**, 2304.

[37] Smoot, M., Ono, K., Ruscheinski, J., Wang, P., and Ideker, T. (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, **27**, 431–432.

[38] Vogel, T., Simonet, P., Jansson, J., Hirsch, P., and Tiedje, J., et al. (2009) TerraGenome: a consortium for the sequencing of a soil metagenome. *Nature Reviews Microbiology*, **7**, 252.

Bertrand Clarke
Department of Statistics
University of Nebraska Lincoln
USA
E-mail address: bclarke3@unl.edu

Camilo Valdes
Center for Computational Sciences
University of Miami
USA
E-mail address: cvaldes3@med.miami.edu

Adrian Dobra
Department of Statistics
University of Washington
USA
E-mail address: adobra@uwashington.edu

Jennifer Clarke
Department of Food Science and Technology
University of Nebraska Lincoln
USA
E-mail address: jclarke3@unl.edu