Joint Analysis of Multiple Metagenomic Samples

Yael Baran[1], **Eran Halperin**[2,3]*

[1] School of Computer Science, Tel-Aviv University, Tel-Aviv, Israel, [2] School of Computer Science and the Department of Molecular Microbiology and Biotechnology, Tel-Aviv University, Tel-Aviv, Israel, [3] International Computer Science Institute, Berkeley, California, Unites States of America

Abstract

The availability of metagenomic sequencing data, generated by sequencing DNA pooled from multiple microbes living jointly, has increased sharply in the last few years with developments in sequencing technology. Characterizing the contents of metagenomic samples is a challenging task, which has been extensively attempted by both supervised and unsupervised techniques, each with its own limitations. Common to practically all the methods is the processing of single samples only; when multiple samples are sequenced, each is analyzed separately and the results are combined. In this paper we propose to perform a combined analysis of a set of samples in order to obtain a better characterization of each of the samples, and provide two applications of this principle. First, we use an unsupervised probabilistic mixture model to infer hidden components shared across metagenomic samples. We incorporate the model in a novel framework for studying association of microbial sequence elements with phenotypes, analogous to the genome-wide association studies performed on human genomes: We demonstrate that stratification may result in false discoveries of such associations, and that the components inferred by the model can be used to correct for this stratification. Second, we propose a novel read clustering (also termed "binning") algorithm which operates on multiple samples simultaneously, leveraging on the assumption that the different samples contain the same microbial species, possibly in different proportions. We show that integrating information across multiple samples yields more precise binning on each of the samples. Moreover, for both applications we demonstrate that given a fixed depth of coverage, the average per-sample performance generally increases with the number of sequenced samples as long as the per-sample coverage is high enough.

Citation: Baran Y, Halperin E (2012) Joint Analysis of Multiple Metagenomic Samples. PLoS Comput Biol 8(2): e1002373. doi:10.1371/journal.pcbi.1002373

Editor: Chris P. Ponting, University of Oxford, United Kingdom

Received August 5, 2011; **Accepted** December 20, 2011; **Published** February 16, 2012

Funding: E.H. is a faculty fellow of the Edmond J. Safra Bioinformatics program at Tel-Aviv University. Y.B. was supported by the Edmond J. Safra fellowship. E.H. and Y.B were also supported by the Israel Science Foundation grant no. 04514831. Y.B was also supported by the IBM Open Collaborative Research Program. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: heran@post.tau.ac.il

Introduction

Metagenomic samples are pooled samples of the genomes of multiple microorganisms living in the same environment. They can be taken either from the outer environment or from microbial populations colonizing other living organisms. Metagenomic studies focus on the taxonomic and functional characterization of the microbial populations contained in such samples. These studies have been boosted by advances in Next Generation Sequencing (NGS) technologies. Particularly, Whole Genome Shotgun (WGS) sequencing provides reads sampled randomly along the genomes, and enables simultaneous phylogenetic and functional analysis of the samples. Although WGS datasets contain plenty of information, they are hard to decipher, as we will further explain below. In a nutshell, the natural way to explore their composition is by aligning the sequencing reads against known databases of whole genomes or of marker genes, however these databases are seriously limited and biased. In addition, one cannot a-priori tell which reads originated from the same genome, and therefore many methods attempt to cluster the reads according to species of origin as a preliminary stage; unsupervised binning methods face an especially hard challenge, and are currently practiced mostly on extremely simple or simulated datasets.

Along with the increasing availability of single-metagenome WGS datasets, datasets consisting of multiple metagenomic samples are also becoming abundant. These datasets typically include samples taken from similar environments, such as ocean water sampled from different locations or depths [1], or microbiomic samples taken from a group of human individuals [2]. To date, the primary analysis of the resulting sequences is performed separately for each sample. Our principal observation is that combining information from multiple samples improves the characterization of each of the samples. We give two demonstrations of this principle: First, we present a method for the unsupervised characterization and quantification of shared hidden components across samples. Second, we present a binning method that operates on multiple samples simultaneously in order to achieve high per-sample precision.

We consider an unsupervised learning approach, in which we aim at learning the shared components of the different samples in an attempt to answer the prominent question of metagenomics, "what's in the mix", without relying on any prior knowledge. While the use of stored sequences of whole genomes [3] or of marker genes, such as the 16S rRNA subunit [4], is currently the most effective way of analyzing large-scale metagenomic samples, it is considerably hindered by the incompleteness of existing databases: In addition to including only a small fraction of the species expected to be found in the samples, the set of species which these databases do include is highly biased, and this bias in turn causes a bias in the analysis results. Supervised analyses also often assume that the properties of the samples which are of biological or medical interest correspond to known taxonomic or

Author Summary

Microorganisms are extremely abundant and diverse, and occupy almost every habitat on earth. Most of these habitats contain a complex mixture of many different microorganisms, and the characterization of these metagenomic mixtures, in terms of both taxonomy and function, is of great interest to science and medicine. Current sequencing technologies produce large numbers of short DNA reads copied from the genomes of a metagenomic sample, which can be used to obtain a high resolution characterization of such samples. However, the analysis of such data is complicated by the fact that one cannot tell which sequencing reads originated from the same genome. We show that the joint analysis of multiple metagenomic samples, which takes advantage of the fact that the samples share common microbial types, achieves better single-sample characterization compared to the current analysis methods that operate on single samples only. We demonstrate how this approach can be used to infer microbial components without the use of external sequence data, and to cluster sequencing reads according to their species of origin. In both cases we show that the joint analysis enhances the average single-sample performance, thus providing better sample characterization.

functional annotations, although this is not necessarily the case. An intriguing counter-example is the recently discovered *enterotype* classes [5], which are three robust classes to which human gut metagenomic samples can be classified. Although generated using a supervised technique, these classes are characterized by a complex combination of the abundance of many bacterial species, which do not correspond to specific taxonomic units.

Aiming to avoid these disadvantages, we developed a method for the inference of hidden components within the data, which leverages on the fact that these unknown components are shared by the different samples. Each of the components is characterized by its sequence composition pattern, specifically the frequency of different short k-mer words in the sequence, which is known to characterize bacteria at different phylogenetic scales [6,7]. Due to the unsupervised nature of the method, we do not expect the components to represent any easily-interpretable biological entity, but instead to provide a composite characterization of the samples. Unlike the enterotypes clustering procedure, our method does not require an alignment stage and does not classify the samples to distinct classes. Instead, we search for the best components that explain the data, and each sample is assigned a distribution over these components; this is done by utilizing Probabilistic Latent Semantic Analysis (PLSA) [8], a technique applied to fields such as information retrieval and natural language processing. Despite these differences, there are some correlations between the inferred components and the enterotypes, which we mention in the Discussion.

Unsupervised component estimation can be used for multiple purposes, and we choose to demonstrate its applicability to a new paradigm for studying statistical association between metagenomic content and phenotypes, which we now introduce. We look for DNA words - long k-mers - whose abundance in the sequencing reads of the different samples correlate with the phenotype at question. For large enough k, differences in the abundance of certain k-mers would capture differences in the abundance of specific species, genes or functional domains which cause the phenotype or are affected by it.

The proposed framework is analogous to the widely used paradigm of Genome Wide Association Studies (GWAS), which is used to test for associations between genetic variants in the human genome and phenotypes. In a typical GWAS the frequency of millions of variants, spanning the entire genome, is compared between a group of cases and a group of controls, and variants whose frequencies differ significantly between the two are considered to be statistically associated with the condition. In the context of metagenomic association, the k-mers are analogous to the genetic variants studied in GWAS, and in both applications the goal is to find statistically significant associations between the measured variants and the condition. However, while GWAS searches for specific mutations which are associated with increased risk for the condition, we aim to capture modifications in the bacterial composition - functional or taxonomic - which are associated with the disease. As in the case of GWAS, the advantages of our approach are its computational efficiency, statistical rigor, cross-study comparability, and the fact that it does not require a supervised stage or comparison to existing databases.

Interestingly, when testing this approach on a publicly available dataset [2] containing 124 deeply sequenced samples of human gut microbiomes collected as part of the MetaHIT (Metagenomics of the Human Intestinal Tract) project, we found that the abundance of a large fraction of the k-mers vary with some of the phenotypes, even for k as small as 3. In the GWAS context this is known as a case of stratification: the null hypothesis of equal distribution between phenotype groups does not hold for the typical variant. For example, when the case and control groups have different ethnic composition, the minor allele frequency of an exceptionally large number of Single Nucleotide Polymorphisms (SNPs) may appear correlated with the disease, but these correlations reflect the fundamental genetic difference between the groups, instead of being relevant to the disease.

In order to correct for the stratification and conduct a proper association analysis, we integrate into the association test the estimates provided by the probabilistic model, specifically the estimated proportion of each component within each sample. We chose to characterize the components according to the short k-mers distribution in the samples. Recently, Meinicke et al. [9] propose to model the k-mers distribution of a single metagenomic sample as a mixture over the distributions of already-sequenced genomes; however, the use of multiple samples in our method allows our method to remain unsupervised.

As a second demonstration of the joint analysis approach we consider the task of binning sequence reads into an unknown set of species. Binning is an important preliminary step for further metagenomic analysis, and has been heavily investigated in the past few years, including the development of multiple unsupervised methods [10–16]; however, all existing methods operate on single samples only. We suggest an unsupervised coverage-based approach, and demonstrate that when the samples share a common species core, information can be integrated between them to improve binning precision. In other words, if one wishes to bin a given sample, then the simultaneous binning of other samples would yield better precision for the original sample. Moreover, we show that for a fixed depth of coverage, dividing the sequencing reads between additional related samples improves precision on the sample of interest.

Methods

Association test for metagenomes

Over the last few years, there have been many reports of associations between the content of metagenomic samples, especially human microbiomic samples, and phenotypes. Different studies report associations with different properties of the samples, such as the abundance of certain taxonomic units, mostly phyla and species, the overall taxonomic and functional diversity of the samples, and the abundance of certain genes or groups of genes, such as those participating in specific metabolic pathways (see Turnbaugh et al. [17] for a comprehensive study of obesity which tested most of these properties). In addition, dimensionality reduction techniques such as PCA [18] are often used on top of the raw data. While examining many properties of the samples allows to capture a wide range of associations, it is not always possible to accumulate results over different studies, in order to perform meta-analysis. In addition, it is hard to perform a rigorous statistical analysis, and especially to control for multiple hypotheses, when many different types of tests are carried out.

As a more rigorous approach, we propose to test the abundance of all possible DNA words of a fixed length for association with the phenotype. This test examines a limited but well-defined group of variants, and hence while it is not expected to capture the entire spectrum of possible associations, its results are statistically robust and easy to compare across studies and accumulate for future meta-analysis studies. It does not require alignment or comparison against any existing database, and therefore it can capture associations with unannotated sequences; due to the latter, the test is also fast and easy to implement.

Formally, for a given value of k, the number of occurrences of all k-mers across each of the samples are normalized to obtain sample-specific relative abundances. The counts of complementary k-mers are summed together as they are indistinguishable in the sequencing data. We denote by x_{ia} the relative abundance of k-mer a in sample i, and by y_i the phenotype of sample i. We test the association between the k-mer and the phenotype by fitting a regression model of the form

$$y_i = c \cdot x_{ia} + d \qquad (1)$$

For a given phenotype y we solve the model for each k-mer a by generating the appropriate vector $(x_{1a} \dots x_{Na})$, where N is the number of samples. We use simple regression and logistic regression for continuous and dichotomous phenotypes respectively.

In the Results section we report that some phenotypes are correlated with a large fraction of the k-mers. These correlations reflect large-scale differences in the genetic composition of the samples between the phenotype groups; specifically, a plausible assumption is that there exists a group of common microbial components, and that each sample is a mixture of these components, in unique proportions. The components might be, for example, different bacterial phyla, and a certain phenotype might correlate with a higher proportion of a certain phylum; since there are differences in sequence composition between the phyla, this would cause phenotype-correlated differences in the distributions of many k-mers. However, we are interested not in the large-scale variation, but in the k-mers which remain correlated with the phenotype after taking this variation into account. Assuming there are B components and denoting by p_{ib} the proportion of component b in sample i, $\sum_{b=1}^{B} p_{ib} = 1$, the estimation of the matrix $\mathbf{P} = p_{ib}$ would allow us to construct a corrected model:

$$y_i = c_0 \cdot x_{ia} + \sum_{b=1}^{B-1} c_b \cdot p_{ib} + d \qquad (2)$$

This equation is similar to equation 1 but includes the additional confounding components as covariates. For a given phenotype y, we again solve the model for each k-mer while keeping the covariate expressions fixed. Due to this addition, the association of a k-mer whose association with the phenotype is explained by the covariates will not be statistically significant, as desired. Note that p_{iB} is not included in the equation because of the linear dependency $p_{iB} = 1 - \sum_{b=1}^{B-1} p_{ib}$.

Probabilistic model for stratification

In order to estimate $\mathbf{P}$ we use the following probabilistic model. We assume that the sequencing reads for N metagenomic samples are given, and that the DNA content of the samples is composed of a common set of components; each read has originated from one of the components, and each component is characterized by a typical distribution over the group of all possible k-mers in the sequence, for some fixed small value of k (e.g., $k = 4$). The model is parametrized by two row-stochastic matrices, $\mathbf{P}_{\mathbf{N} \times \mathbf{B}}$ and $\mathbf{F}_{\mathbf{B} \times |\mathcal{K}|}$: the ith row of $\mathbf{P}$, denoted p_{i*}, defines a sample-specific multinomial distribution over all components, and the bth row of $\mathbf{F}$, denoted f_{b*}, defines a component-specific multinomial distribution over $\mathcal{K}$, the group of all k-mers. When we sample a short k-mer from a random position on the reads of sample i, we first sample a component b according to the distribution p_{i*}, and then sample the k-mer according to the distribution f_{b*}. Being defined as general multinomial distribution, some entries in $\mathbf{P}$ and $\mathbf{F}$ may have a zero value; in particular, some components might not be represented in some of the samples.

We note that while the previous subsection discussed long k-mers (e.g., $k \geq 8$), which are each tested for association with the phenotypes, in this section we use short k-mers as characteristics of the components we attempt to learn. Specifically, we chose to use $k = 4$ since it has been shown that 4-mer distributions are characteristic of phylogenetic units [6,7], and since the 4-mer distribution captures both the codon distribution and possible codon biases.

We now turn to calculate the likelihood of the metagenomic data. Since the model explains the k-mer distribution in the reads, we extract the first k-mer from each read, and denote by n_{ia} the number of times k-mer a was extracted from sample i. The likelihood of the counts data R is

$$\mathcal{L}(\mathbf{F} = f, \mathbf{P} = p; R) = \prod_{i=1}^{N} \prod_{a \in \mathcal{K}} \Big(\sum_{b=1}^{B} p_{ib} f_{ba} \Big)^{n_{ia}} \qquad (3)$$

where N is the number of samples, B is the number of components, and $\mathcal{K}$ is the group of all possible k-mers, $|\mathcal{K}| = \frac{4^k + 4^{\lceil \frac{k}{2} \rceil}}{2}$ as the counts for complementary k-mers are joined. Our goal is to estimate the distributions in the matrix $\mathbf{P}$, which are the distributions over the components for each sample.

We note that there is a simple relation between x_{ia} and the above notation, given by $x_{ia} = \frac{n_{ia}}{\sum_{a' \in \mathcal{K}} n_{ia'}}$. Furthermore, under the model assumptions, we have that $E[x_{ia}] = \sum_{b=1}^{B} p_{ib} f_{ba}$. One can verify that if there is a solution such that $\sum_{b=1}^{B} p_{ib} f_{ba} = x_{ia}$, then this solution maximizes the likelihood in Equation 3. Thus, we can view the maximization of the likelihood function as an approximation of the factorization of the $\mathbf{X} = x_{ia}$ matrix, which is row-stochastic, into two other row-stochastic matrices:

$$\mathbf{X}_{N \times |\mathcal{K}|} = \mathbf{P}_{N \times B} * \mathbf{F}_{B \times |\mathcal{K}|}$$

This factorization corresponds to a set of $N \times |\mathcal{K}| + N + B$ linear equations, to which each additional sample adds B variables but also a much larger number of equations - $|\mathcal{K}| + 1$; this could serve as an intuition for the advantage conveyed by sample multiplicity. In addition, this factorization is a variant of the non-negative matrix factorization (NMF) technique, with the added stochasticity constraints (so that the sum of each row in $\mathbf{P}$ and $\mathbf{F}$ is 1). NMF is used to unveil hidden structures within data, and its major advantage over methods such as the widely-used PCA is the high interpretability of the inferred components [19]. A recent paper [20] used NMF in a metagenomic context, however the factorized data matrix was generated using alignment to sequenced genomes, in contrast to our method which does not rely on prior knowledge. The stochasticity constraints turn our model to an exact instantiation of PLSA [8,21], a generative model from the statistical literature. PLSA was originally applied to the field of text analysis for the discovery of topics in a corpus of documents [22]. Due to its great flexibility, it was successfully applied to multiple problems in the field of text learning [23–25] as well as to image content analysis tasks [26].While strong similarities exist between PLSA and NMF, the fact that PLSA is based on a probabilistic model allows us to refine the model to better match the properties of the sequencing data, as we do below.

In the above procedure we extract only the first k-mer from each read because the model assumes that the k-mers are sampled independently according to $\mathbf{F}$, conditioned on the read's component. Extracting multiple k-mers would result in a deviation from the model due to the dependencies between neighboring k-mers on the same read, as well as dependencies between k-mers extracted from multiple reads covering the same genomic region.

Interestingly, the simulations presented in the Results section demonstrate that extracting multiple k-mers from the same read improves performance, despite the dependencies. The reason is that under reasonable coverage and when k is not too large, the relative abundances x_{ia} approach a constant value, and the exact sampling strategy has no effect on the final counts data. It is therefore of benefit to extract multiple k-mers from each read when processing the sequencing reads, however it turns out that the best strategy is to choose not all k-mers present on the read but only a subset, while using a slightly different model, as we explain next.

A refined model. Once multiple k-mers from the same read are used, the model can be refined to reflect the fact that all k-mers extracted from the same read were sampled from the same component. The likelihood function below entails this information:

$$\mathcal{L}(\mathbf{P}=p,\mathbf{F}=f;R) = \prod_{i=1}^{N} \prod_{r \in R_i} \sum_{b=1}^{B} p_{ib} \prod_{a \in K_r} f_{ba} \tag{4}$$

where R_i is the set of all reads sampled from sample i, and K_r is the multiset of the k-mers extracted from read r.

The refined model, like the previous model, assumes that the k-mers on a read are sampled independently given the component, but unlike the previous model, the refined model is sensitive to deviations from this assumption. The reason is that the refined model assigns each read to a component, as opposed to each k-mer, and is therefore more prone to local distortions in the k-mer distribution which result from the dependencies. As we show in the Results section, an effective way to alleviate this problem is to sample a smaller number of k-mers which are more sparsely dispersed along the read. As a result, sampling a relatively small number of k-mers from each read and using the refined model turns out to be an effective strategy for using the sequencing reads in order to estimate $\mathbf{P}$.

Parameter estimation using EM

We use Expectation-Maximization algorithms in order to approximate the maximum likelihood solutions of both the original model (defined in Equation 3) and the refined model (Equation 4), beginning with the refined. The observed variables are groups of extracted k-mers, one group for each read, and the latent variables are the assignments of a component to each of the reads. Let M be the unknown assignments, and let n_{ra} be the number of occurrences of k-mer $a \in \mathcal{K}$ in the multiset of k-mers extracted from read $r \in R$, denoted K_r. The algorithm can now be written as

E-step:

$$\begin{aligned}
Q(p,f|p^{(t)},f^{(t)}) &= E_{M|R,p^{(t)},q^{(t)}}[\log \mathcal{L}(\mathbf{P}=p,\mathbf{F}=f;R)] \\
&= E_{M|R,p^{(t)},f^{(t)}}\left[\sum_{i=1}^{N}\sum_{r \in R_i} \log\left(p_{iM(r)} \prod_{a \in K_r} f_{M(r)a}\right)\right] \\
&= E_{M|R,p^{(t)},f^{(t)}}\left[\sum_{i=1}^{N}\sum_{r \in R_i} \left(\log p_{iM(r)} + \sum_{a \in K_r} \log f_{M(r)a}\right)\right] \\
&= \sum_{i=1}^{N}\sum_{b=1}^{B} C_{ib} \log(p_{ib}) + \sum_{b=1}^{B}\sum_{a \in \mathcal{K}} D_{ba} \log(f_{ba})
\end{aligned}$$

where

$$C_{ib} = \sum_{r \in R_i} \frac{p_{ib}^{(t)} \prod_{a \in K_r} f_{ba}^{(t)}}{\sum_{b'=1}^{B} p_{ib'}^{(t)} \prod_{a \in K_r} f_{b'a}^{(t)}}$$

$$D_{ba} = \sum_{i=1}^{N}\sum_{r \in R_i} n_{ra} \frac{p_{ib}^{(t)} \prod_{a' \in K_r} f_{ba'}^{(t)}}{\sum_{b'=1}^{B} p_{ib'}^{(t)} \prod_{a' \in K_r} f_{b'a'}^{(t)}}$$

M-step:

$$p_{ib}^{(t+1)} = \frac{C_{ib}}{\sum_{b'=1}^{B} C_{ib'}}$$

$$f_{ba}^{(t+1)} = \frac{D_{ba}}{\sum_{a' \in \mathcal{K}} D_{ba'}}$$

The running time of each iteration of the above algorithm is $O(B|R|L)$, L being the number of k-mers extracted from each read. For realistic values of $|R|$ this is time consuming. In addition, for large datasets the entire data cannot fit in memory. Consider, for example, the case where the number of individuals is $N = 100$, the number of reads per individual is $3 \cdot 10^7$, and all non-overlapping 4-mers from reads of length 80 bp are used. In this case, the amount of memory required is at least 60 GB, even if every nucleotide letter is stored in two bits. We note that by changing the order of the summation, one can use a considerably smaller amount of memory, however in each iteration of the EM the entire dataset will have to be read again to memory. Therefore, when analyzing large datasets we recommend to use the simpler model, described by Equation 3, which ignores the relation between k-mers on the same read. The input counts n_{ia} are extracted in a single pass through the data. The latent variables M are the assignments of each pair (sample, k-mer) to a component, and the EM algorithm is as follows:

E-step:

$$Q(p,f|p^{(t)},f^{(t)}) = E_{M|R,p^{(t)},f^{(t)}}[\log \mathcal{L}(\mathbf{P}=p,\mathbf{F}=f;R)]$$

$$= \sum_{i=1}^{N}\sum_{b=1}^{B} C_{ib}\log(p_{ib}) + \sum_{b=1}^{B}\sum_{a\in\mathcal{K}} D_{ba}\log(f_{ba})$$

where

$$C_{ib} = \sum_{a\in\mathcal{K}} \frac{n_{ia}p_{ib}^{(t)}f_{ba}^{(t)}}{\sum_{b'=1}^{B} p_{ib'}^{(t)}f_{b'a}^{(t)}}$$

$$D_{ba} = \sum_{i=1}^{N} \frac{n_{ia}p_{ib}^{(t)}f_{ba}^{(t)}}{\sum_{b'=1}^{B} p_{ib'}^{(t)}f_{b'a}^{(t)}}$$

The M-step is similar to the one described for the refined model. Note that the running time of each iteration of the EM algorithm is now $O(NB4^k)$, and it is therefore very efficient as long as k is fixed.

Binning algorithm

The model we presented infers common components in the samples but does not assign the reads to these components; it provides for each read a probability distribution over the components that could be used for an assignment, but in general is not optimized for this goal. Such assignment, or binning, is an important preliminary step in the analysis of metagenomic samples, especially binning according to species of origin. We therefore devised an unsupervised algorithm which performs binning over multiple samples simultaneously, again leveraging on sample similarity, this time assuming a common species core. Most previous unsupervised binning methods are based on sequence composition [10–15]. For example, CompostBin [11] computes for each read its 6-mer distribution, similarly to the process performed by our component inference algorithm, and clusters these distributions using spectral methods. The main limitation of composition-based approaches is that they require relatively long reads (1000 bp in the case of CompostBin) due to the variance in sequence properties along the genome. Recently, a coverage-based method, AbundanceBin, was developed [16] with the advantage of being able to bin even very short reads (as small as 75 bp). Since it relies on abundance differences for binning, AbundanceBin is only able to discern between species whose abundance levels are considerably different (they report that a ratio of 2:1 is required). Our algorithm is also coverage-based, but because it operates on multiple samples it can use abundance difference in any of the samples to tell between such species.

Assume we are given N metagenomic sequencing samples, consisting of a total of b bacterial species. We wish to divide the reads in all samples into b bins that correspond to the species from which they were sequenced. The binning algorithm, which we term MultiBin, proceeds as follows:

1. Pool the reads from all the samples together, and perform pairwise alignment between all pairs. For each pair, check whether the alignment shows a long overlap between the two sequence reads, suggesting that the two reads originated from the same genome. Put differently, we generate a graph $G=(V,E)$, where the set V corresponds to the reads, and the set of edges E corresponds to pairs of reads with a substantial overlap. In our experiments we demanded an overlap of at least 50 bases per sequence.
2. Greedily find a maximal independent set in G. We call the reads in this set *tags* and denote it by T. Following this process each read r is either a tag, or is affiliated with a single tag which substantially overlaps it, $TAG[r]$. For each tag read t, we denote by c_{ti} the number of reads from sample i which substantially overlap it (but are not necessarily tagged by it), and by v_t the vector $(c_{t1}\dots c_{tN})$.
3. Perform k-medoids clustering on the set of vectors $\{v_t|t\in T\}$, starting from a random choice of b centers. Wait for convergence and divide the tags into bins according to the clustering result. Assign every non-tag read r to the same bin to which $TAG[r]$ was assigned.

In the last stage, the distance between every two vectors c_t, $c_{t'}$ was computed as $\sum_{i=1}^{n}\frac{(c_{ti}-c_{t'i})^2}{c_{ti}+c_{t'i}}$. K-medoids clustering was performed using a local search procedure, in which we start from a random choice of centers and attempt to improve the solution by swapping at least one of the centers with another vertex, until no further improvement can be made.

The initial stage of distance computation takes $O(|T|^2N)$, and each iteration of the clustering algorithm takes $O(|T|^2b)$; in our experiments convergence was reached within three iterations or less. The running time of the alignment stage, as well as the size of T, depends on the composition of the mixtures. We note that clustering is performed only on the tag reads, whose number is approximately bounded by the sum of the genome sizes in the samples divided by the read length.

The above procedure assumes that the number of species in the samples b in known. In the Results section we describe a procedure for determining the number of species based on examining the clustering results for different numbers of bins.

Results

We evaluated our methods using both real data and simulated data. We used the MetaHIT dataset (downloaded from EBI, accession ERA000116), which includes over 0.5 terabases of sequence generated from the gut microbiomes of 124 European individuals using the Illumina Genome Analyzer technology. The average amount of sequence per individual is 4.5 gigabases, and the paired-end read length is 44 or 74, depending on the sample. We used the publicly available raw reads, which were obtained after filtering human and Illumina adapter contaminant reads and low quality reads. The sampled individuals vary on the following variables: country of origin (Denmark/Spain), age, BMI (Body Mass Index), gender, and status for infectious bowl diseases (Ulcerative Colitis/Crohn's disease/disease free). In the context of this paper all the variables will be referred to as phenotypes of the human host, although country and age are in fact determinants of the metagenomic content, instead of being affected by it.

Large differences in k-mer distributions between phenotype groups

In the initial stage we simply compared, for each k-mer, its relative abundance in different phenotype groups using a two-sample *t*-test. Surprisingly, the relative frequencies of many of the k-mers are significantly correlated with many of the phenotypes. This is true even for k as small as 3: for example, the frequency of 69% of the 3-mers and of 61% of the 4-mers differs between the Spanish and the Danish samples at the 0.05 level. To the best of our knowledge, such dramatic differences in sequence composition between samples from different countries have not been observed previously. This effect might be partially due to differences in sample preparation and DNA extraction procedures, which are known to exist between the MetaHIT samples from the two countries; however, we also observed significant differences across phenotypes within each country: The frequency of 40% of the 4-mers differs between the Crohn and the healthy Spanish samples, and the frequency of 18% of the 4-mers differs between the 10 highest- and lowest-BMI Danish groups. Permuting the phenotype labels 10^5 times yielded p-values of 0.0027, 0.0521 respectively for these fractions of rejected nulls.

A possible concern regarding the counts statistics are possible biases in the GC content distribution of the reads. We note that unlike different single-genome samples, different metagenomic samples are not expected to contain the same sequence composition characteristics, and therefore normalizing for such biases is a challenging task. We note that in the context of an association study between a phenotype and the metagenome, it is possible to avoid this problem using a permutation test, at the expense of power reduction.

Multi-sample modeling of bacterial stratification

The majority of the correlations between k-mers and phenotypes are false positives resulting from a hidden stratification which confounds the k-mer distributions. In order to reveal the components and to quantify them, we solved the probabilistic model described in the Methods section. Because the MetaHIT dataset is large, we used the more efficient version of the algorithm (which solves Equation 3). The input to the algorithm is a counts matrix of size $[124 \times 136]$, detailing for each of the 124 samples how many occurrences of each possible 4-mer it includes (there are only 136 possible 4-mers, instead of $4^4 = 256$, because complementary strings are indistinguishable in the sequencing data). Since extracting multiple 4-mers from each read did not seem to considerably change the results in this particular case (this does not hold in general, as illustrated later), we used only the first, highest-quality k-mer. The model was solved for three components ($B = 3$) by running EM multiple times from random starting points and choosing $\hat{P}$ from the maximum-likelihood run. The solution $\hat{P}$ provides, for each sample i, the components proportions p_{i1}, p_{i2} and p_{i3}, such that $\forall i \sum_{b=1}^{3} p_{ib} = 1$.

We first tested each of the components $b = 1,2,3$ for correlation with each of the phenotypes in the following regression model:

$$y_i = c \cdot p_{ib} + d \tag{5}$$

We solved this model for each phenotype y and for each cluster b, attempting to discover biologically meaningful components. Simple regression and logistic regression are used for continuous and dichotomous phenotypes, respectively. As can be seen in Table 1, highly significant p-values were obtained for predicting country among healthy individuals and for predicting BMI among the Danish. These results remain consistent also after correcting for the other measured phenotypes by entering them as covariates into the regression models. Results significant at the 0.05 level were obtained also for predicting colitis and Crohn status among the Spanish. In the case of Crohn's disease, the power of the regression model was limited due to the small number of cases (only 4), but 10^4 permutations yielded a p-value of $7.3 \cdot 10^{-4}$. Figure 1 visually demonstrates the separation of Crohn cases and controls on the plane defined by the components proportions.

We note that we also solved the model for $B = 4,5,6,7$. In these cases we found that the smallest per-component p-values, as well as the proportions of explained phenotypic variance captured by all components together, were similar to those obtained for $B = 3$. We therefore report the estimates for three components throughout the paper.

The components estimates correct for stratification

The results from the last section, showing that the components proportions are correlated with some of the phenotypes, suggest they indeed may be used to correct the association between these phenotypes and long k-mers; we therefore attempted to perform this correction on 8-mers.

Figure 2 demonstrates such a successful correction. Two quantile-quantile curves compare the uniform distribution to the distribution of the p-values obtained by testing association between all possible 8-mers and BMI within the Danish samples. The black curve shows the uncorrected p-values, and its shape reflects the fact that they are highly deflated. The red curve shows the p-values obtained by adding the components proportions to the regression equation; this curve approaches the identity, indicating that these statistics capture the variance in the phenotype explained by most of the 8-mers.

A comparison between the likelihood models

We compared the precision of the two models, the original and the refined, utilizing different strategies for k-mer extraction. Performance was tested on simulated data, each simulation consisting of 100 mixtures of the following 4 bacterial species: *Listeria monocytogenes* (phylum Firmicutes), *Bacteroides vulgatus* (phylum Bacteroidetes), *Bifidobacterium longum* (phylum Actinobacteria),

Table 1. Significant correlations between components proportions and phenotypes.

predicted variable	component 1	component 2	component 3
country within healthy	$9.2 \cdot 10^{-5}$ $(1.0 \cdot 10^{-7})$	$5.5 \cdot 10^{-3}$ $(2.8 \cdot 10^{-3})$	$2.6 \cdot 10^{-1}$ $(2.5 \cdot 10^{-1})$
BMI within Denmark	$9.4 \cdot 10^{-3}$ $(8.8 \cdot 10^{-3})$	$6.3 \cdot 10^{-1}$ $(6.2 \cdot 10^{-1})$	$7.6 \cdot 10^{-3}$ $(7.0 \cdot 10^{-3})$
Colitis within Spain	$3.6 \cdot 10^{-2}$ $(2.3 \cdot 10^{-2})$	$1.6 \cdot 10^{-1}$ $(1.5 \cdot 10^{-1})$	$2.0 \cdot 10^{-1}$ $(4.0 \cdot 10^{-1})$
Crohn within Spain	$1.5 \cdot 10^{-2}$ $(7.3 \cdot 10^{-4})$	$2.9 \cdot 10^{-1}$ $(2.9 \cdot 10^{-1})$	$3.3 \cdot 10^{-2}$ $(8.7 \cdot 10^{-3})$

The predicted variables were regressed on the proportions of each component separately. The table gives the regression p-values, and in parentheses the empirical p-values obtained by permuting the components proportions 10^7 times while keeping the phenotypes constant.

doi:10.1371/journal.pcbi.1002373.t001

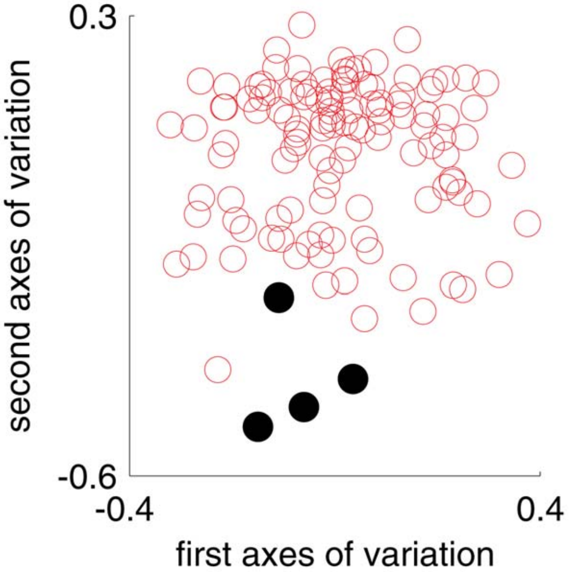


Figure 1. Crohn's disease status separates with components proportions. Each marker corresponds to an individual, red for Crohn-free and filled black for Crohn cases. The markers are positioned on the two-dimensional plane defined by the components proportions (there are three components but only two dimensions because the proportions sum to 1). The Crohn-free individual at the bottom part of the figure is a colitis case.
doi:10.1371/journal.pcbi.1002373.g001

Pseudomonas stutzeri (phylum Proteobacteria). The components distributions matrix $\mathbf{P}$ was randomly drawn from the uniform distribution and normalized to row-stochastic, and the k-mer distributions matrix $\mathbf{F}$ was computed from the actual genomes. For each sample 1,000 sequencing reads of length 100 bp were simulated by sampling a bacterium according to P, and then a random position in the genome of that bacterium as the starting point of the read. We chose a read length of 100 bp since it is currently a length that is obtained by most high-throughput sequencing technologies.

Figure 3 compares the effect of extracting different subgroups of k-mers along each read and using them either in the original or in the refined model. Under the original model, extracting multiple k-mers improves estimation precision of $\mathbf{P}$ compared with extracting only the first k-mer; this is the case even when these k-mers overlap, and are therefore highly correlated. Shifting to the refined model greatly improves the estimation precision when choosing all non-overlapping k-mers along the read; however, even better results are obtained when using only nine sparsely dispersed k-mers along the read, as using a small number of distant k-mers decreases the dependencies of k-mer sequence between and within reads.

A comparison between PLSA and PCA

PLSA is a dimensionality reduction method, and we therefore compared its performance to the widely used principal component analysis (PCA). PCA has been extensively used in metagenomic studies [2,5] for sample visualization and classification. We used the simplified setting in which the counts matrix n is generated from mixtures of multinomials: Setting $N=100, B=3, |\mathcal{K}|=256$ we generated $\mathbf{P}$ and $\mathbf{F}$ by drawing their entries from the uniform distribution followed by normalizing to row-stochastic, and then generated the counts matrix n by sampling each row n_{i*} according to the multinomial distribution specified in $(\mathbf{P}*\mathbf{F})_{i*}$, with varying numbers of counts per sample.

Both PLSA and PCA were tested in the task of estimating the matrix $\mathbf{P}$. Since PCA operates with no stochasticity constraints, we estimates precision as the average squared correlation coefficient (r^2) between the true vectors p_{*1}, p_{*2}, p_{*3} and either the three strongest principal components (for PCA) or the vectors $\hat{p}_{*1}, \hat{p}_{*2}, \hat{p}_{*3}$ (for PLSA). For both methods we chose the ordering of the components that yielded the highest score.

Figure 4 shows that PLSA's estimates of $\mathbf{P}$ are considerably more accurate than those obtained by PCA. This result confirms that PLSA is indeed a more appropriate method for characterizing mixture components in our context.

Joint binning over multiple samples

Nine datasets, each consisting of five metagenomic samples, were generated using MetaSim [27]. All samples in a given dataset contained the same set of bacterial species in different, randomly drawn proportions. The datasets differed in the number of species they contain, ranging from 2 species to 10 species in each sample of the most complex dataset. The species distribution of a sample containing $2<k$ species was generated incrementally by adding a random number sampled uniformly from [0,1] to the species proportions of an existing $k-1$ sample and normalizing to 1. In the first experiment we generated 100,000 reads of length 400 bp from each sample, and in the second experiment 400,000 reads of length 75 bp each. Overlaps between reads were determined by running BLAT [28] and requiring an exact match at the edges of the reads; the BLAT parameters we used restricted the results to matches of length ~ 50 bp and above. Precision was computed as the fraction of reads assigned to the correct species, averaged over all species.

We compared the precision obtained by MultiBin to the performance of AbundanceBin, a program implementing the equivalent coverage-based approach which was shown to perform precise binning of species exhibiting different abundance levels using reads of lengths 400 and 75 bp. AbundanceBin operates on single samples only, and therefore was run separately on each sample, while MultiBin was run on all five samples in each dataset simultaneously. We note that the separate execution of AbundanceBin conveys no information about the correspondence between the bins across samples, and so we chose the best matching between the bins and the species in each sample so as to maximize the total precision.

As can be seen in Figure 5, MultiBin performs better than AbundanceBin over both read lengths and over all dataset complexities. For the 400 bp reads MultiBin maintains a precision of over 0.8 even for mixtures of five species. MultiBin is also able to bin the 75 bp reads, although with lesser success; we note that the ability to bin short reads is unique to coverage-based approaches, and that in principle there is no advantage in having longer reads, assuming the coverage is high enough. As for AbundanceBin, its performance on the simple mixtures exhibits high variation between samples because of its reliance on large abundance differences within each sample. AbundanceBin's performance also deteriorates more rapidly as the number of species per sample increases compared with MultiBin.

We went on to evaluate MultiBin under the realistic scenario in which the reads have sequencing errors. To do so, we adjusted the alignment stage of MultiBin, currently performed by BLAT, to be more permissive. We tested this modification by again generating for the above datasets 100,000 reads of length 400 bp, this time introducing base substitutions into the reads. When the substitution rate was increased to 2% and then to 5%, the precision for mixtures of two species decreased from 1.0000 to 0.9975 and then to 0.9966. Overall we conclude that the effect of these errors is not

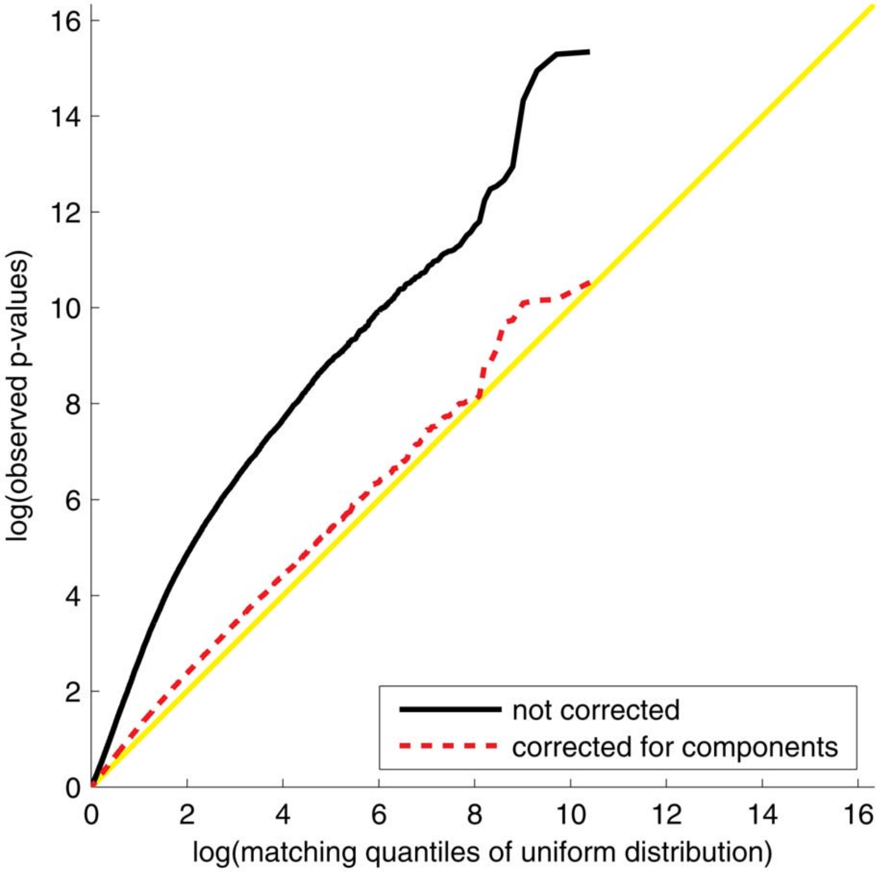


Figure 2. Association between 8-mers relative abundances and BMI can be corrected using the components proportions. Quantile-quantile curves comparing the uniform distribution to the distribution of the p-values for association between all 8-mers and BMI within the Danish samples. The uncorrected p-values are highly deflated (black), indicating that the abundance of many 8-mers is correlated with BMI. However, when the components proportions are added the the regression equation (red), the correlation disappears for most 8-mers.
doi:10.1371/journal.pcbi.1002373.g002

dramatic, and that for realistic error rates they could be largely moderated by adjusting the alignment procedure.

We note that integrating information across samples enables MultiBin to perform precise binning even when the variance in species distribution across the samples is relatively small. For example, when simulating 400 bp reads from five nearly-balanced mixtures of two species - the relative abundance of the more abundant species were 0.57, 0.55, 0.70, 0.53 and 0.56 - AbundanceBin still obtained a precision of 0.93. These results also demonstrate that MultiBin achieves precise binning on nearly-balanced samples; in contrast, a coverage-based method which did not integrate information across samples would produce extremely poor results on each of these samples alone.

Determining the number of clusters is an issue widely explored in the literature, and particularly, several approaches exist and have been tested for similar problems [29]. We found that running the algorithm multiple times for different values of b, measuring the Hartigan index [29,30] for each value and choosing the value at which the index decreases sharply and reaches a plateau gave accurate results, as long as the binning itself was accurate enough (precision of ~ 0.8 and above).

Sample multiplicity in the design phase

Our results so far demonstrate that joint modeling of multiple metagenomic samples can be helpful in the analysis stage. A further question has to do with the design stage: Given a fixed coverage depth and a potential pool of related metagenomic samples, how many of the samples should be sequenced in order to achieve optimal characterization of the underlying microbial composition? There seems to be a tradeoff between sequencing with high coverage a small number of samples and the superficial sequencing of many samples. We tested this tradeoff in both the components estimation problem and the binning task.

For components estimation, our task is to best characterize the components, or in other words to estimate the $\mathbf{F}$ matrix; the $\mathbf{P}$ matrix varies with the number of samples and therefore cannot be compared here. We simulated instances of the proposed probabilistic model by uniformly drawing the P and F matrices followed by normalization to row-stochastic, and then drawing observations from the corresponding multinomial distributions. We used $B=4$ components, each defined by a multinomial distribution over $4^4=256$ possible results. Initially, $\mathbf{P}$ was defined for 1,000 samples, and for each sample 1,000 counts were drawn. The model was then solved for a decreasing number of samples by joining samples together to obtain 1000, 500, 200, 100, 50, 20, 10, 5, 2 and 1 samples. As can be seen in Figure 6(a), the optimal estimation of $\mathbf{F}$ was achieved for 50 samples, each including 20,000 counts: Increasing the number of samples further past this point does not allow enough data to be gathered from each sample, resulting in a decrease in performance.

For the binning task, we ran our algorithm on mixtures of 15 species and reads of length 400 bp. Due to efficiency considerations

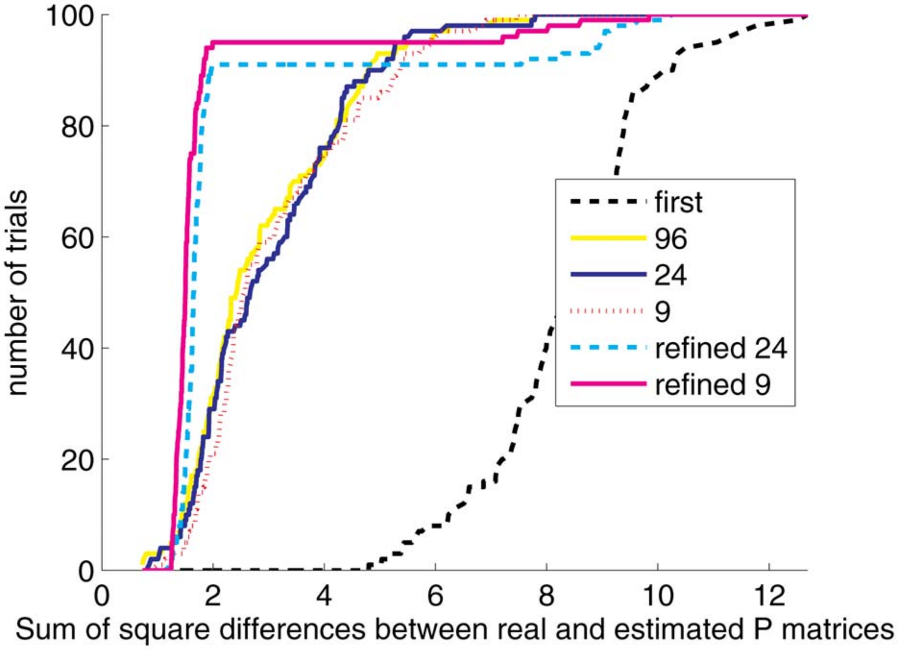


Figure 3. Strategies for *k*-mer choice under the different models. Error is measured as the sum of squared differences between true and estimated **P** matrices. The plot presents, for different error thresholds, the number of runs out of 100 which yielded a precision at least as small as the threshold. Using the original model, we extracted from each 100 bp read either the first *k*-mer, 9 sparely dispersed *k*-mer along it, 24 non-overlapping *k*-mers or 96 overlapping *k*-mers. Extracting multiple *k*-mers can be seen to increase precision considerably. Shifting to the refined model yields an even better precision; since this model is more sensitive to dependencies between *k*-mers, extracting only the fewer dispersed *k*-mers is preferable over extracting all non-overlapping *k*-mers.
doi:10.1371/journal.pcbi.1002373.g003

we did not produce actual reads using MetaSim, but instead drew read start positions randomly along the genomes, and determined an overlap for reads which physically overlapped by more than 100 bp at the edges. Figure 6(b) shows the resulting precision when binning is performed over a fixed number of 8,192,000 reads allocated to an increasing number of samples, in which the species

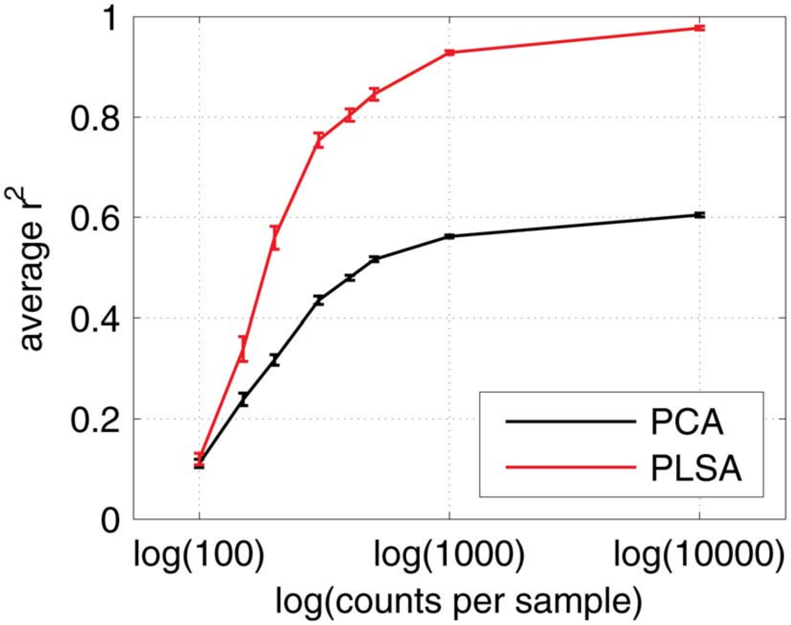
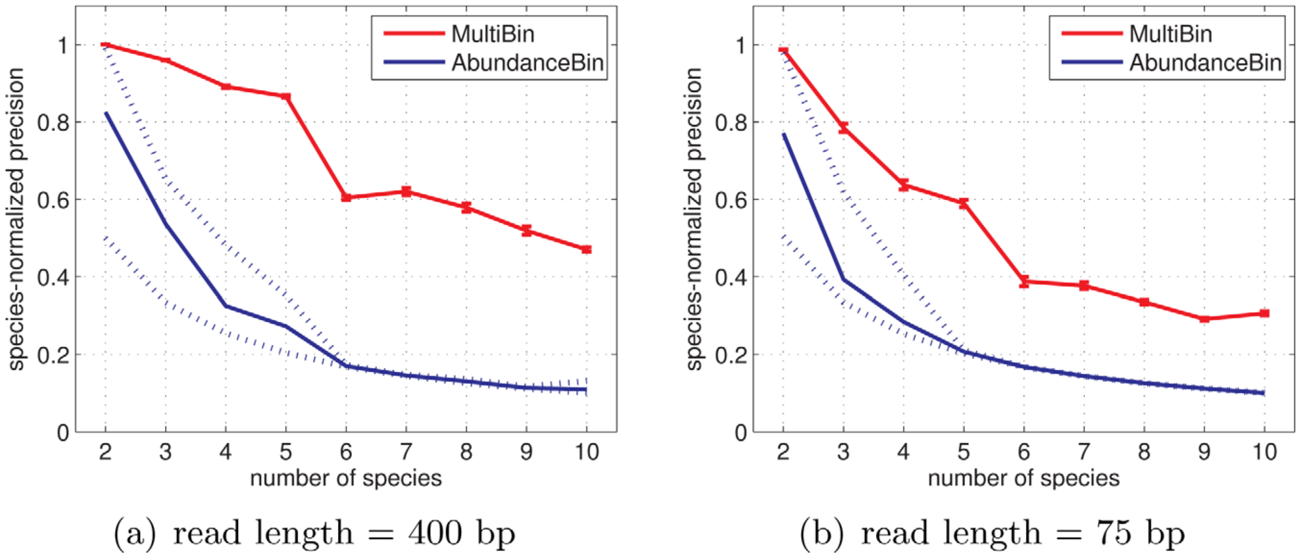


Figure 4. PLSA approximates mixture coefficient better than PCA. PCA and PLSA were performed on a simulated counts matrix n with $N=100, B=3, |\mathcal{K}|=256$ and different number of per-sample counts. The plot shows the average squared correlation coefficient between the true vectors p_{*1}, p_{*2}, p_{*3} and the three strongest principal components (in the case of PCA) or PLSA estimates $\hat{p}_{*1}, \hat{p}_{*2}, \hat{p}_{*3}$. For each per-sample counts value 20 experiments were performed, and the plot gives the mean result and the standard error of the mean. The estimates obtained by PLSA show higher correlation with the true mixture proportions.
doi:10.1371/journal.pcbi.1002373.g004



(a) read length = 400 bp

(b) read length = 75 bp

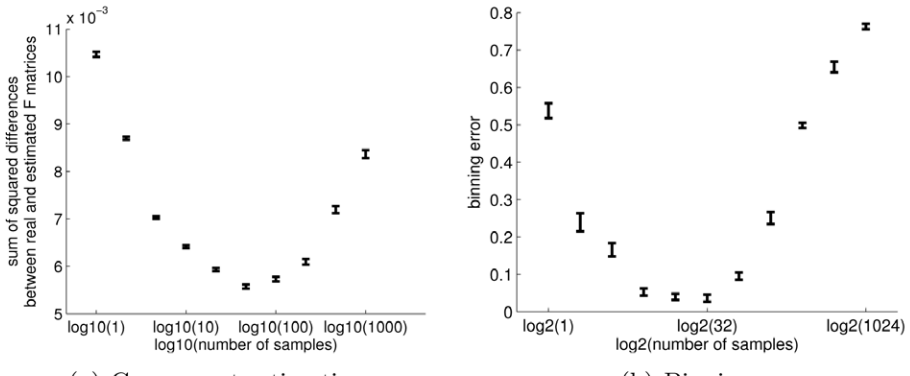
Figure 5. Simultaneous binning over multiple samples achieves higher precision compared with the equivalent single-sample approach. MultiBin and AbundanceBin were both run on datasets of increasing complexity. Each dataset is composed of 5 mixtures of the specified number of species. The specified precision is the proportion of reads correctly assigned to a bin, averaged over all species. For MultiBin (red) the curves show average precision over 10 random starts of the clustering algorithm, and the error bars give the standard error of the mean. For AbundanceBin (blue) the curves show the average precision over the 5 samples in the dataset, and the dashed lines give the highest and lowest result of the 5. MultiBin achieves consistently better precision over both read lengths and over all sample complexities. AbundanceBin's performance exhibits high between-sample variability, and also deteriorate more rapidly as the number of species increase.
doi:10.1371/journal.pcbi.1002373.g005

proportions are again uniformly drawn. The highest precision is obtained for 32 samples.

Both plots demonstrate that sample multiplicity is an advantage given a fixed coverage: as long as the per-sample coverage is reasonable, allocating the sequencing reads to as many samples as possible improves components characterization and binning precision.

Discussion

We demonstrated the advantage in joint modeling of multiple metagenomic samples, by showing that it allows the unsupervised inference of hidden genetic component, and increases the precision of coverage-based binning. This advantage holds for both the analysis and the design stage; as for the latter, the results suggest that when wishing to characterize a given metagenomic sample, it is useful to divide the coverage between additional samples from similar environments. It might also be possible to apply a biological or chemical treatment to some of the samples, which would further accentuate the differences between them; when the samples are analyzed jointly, these differences are expected to further enhance performance. Similarly, sequencing data available from previous experiments can be used to improve the analysis of new samples. A similar tradeoff between the number of samples and per-sample coverage has been observed for testing the power of rare variant



(a) Component estimation

(b) Binning

Figure 6. Increasing the number of samples for a fixed depth of coverage improves both components characterization and binning precision. Left: A fixed number of 10^6 counts were generated from a model defined by uniformly drawn $\mathbf{P}$ and $\mathbf{F}$ matrices using $B=4, |\mathcal{K}|=256$. The value of N, the number of samples, varied from 1 to 1000, and 100 trials were performed for each value. The highest average precision of $\mathbf{F}$ estimation is obtained for $N=50$. **Right:** A fixed number of 8,192,000 reads of length 400 bp were sampled from different numbers of samples, each consisting of 15 species in uniformly drawn proportions. The smallest average error over all samples was obtained when 32 samples are sequenced. In both plots the error bars give the standard error of the mean.
doi:10.1371/journal.pcbi.1002373.g006

discovery in sequencing data [31], and sample multiplicity is likely to become a key issue in the future design of both standard and metagenomic sequencing studies.

We note that the components estimates obtained by solving the probabilistic model are interesting by themselves, outside of the context of association correction; particularly, they can allow for the characterization of variability patterns in metagenomic samples and for sample classification. Different estimates will be obtained by setting the parameters k and B to different values; our choice of $k=4$ and $B=3$ was meant to capture a high-level division, possibly taxonomic, of the microbial population, as it is known that bacterial phyla have characteristic sequence composition. A recent paper [5] identified metagenomic variability components using a supervised approach and divided the Danish samples accordingly to discrete classes termed *enterotypes*; interestingly, there are some correlations between these enterotypes and the components we obtain. For example, samples belonging to the first enterotype have a low proportion of the fourth component (p-value $=1.53 \cdot 10^{-6}$) when solving the model for $k=4$ and $B=4$, and those belonging to the second enterotype have a low proportion of the first component ($6.95 \cdot 10^{-4}$) when solving the model for $k=4$ and $B=3$. However, as explained in the Methods, our algorithm is fundamentally different from the PCA used to identify the enterotypes, and is expected to yield components of different nature, on top of it being unsupervised.

As for the proposed binning algorithm, unlike most other algorithms it can be used on datasets containing short reads, since the reads need only be long enough so as to determine unique sequence overlaps between them. In addition, the algorithm can be further improved to use not only coverage information but also other features, such as sequence composition, by adding them to the vectors on which clustering is performed.

Lastly, the implementation of the proposed association test for the MetaHIT dataset was limited by the sequencing quality, which forced us to extract only the first k-mer from each read and therefore to examine only relatively short k-mers ($k=8$), otherwise the counts data would become too sparse. We believe that this problem could be addressed by the integration of sequencing uncertainties into the counts data. With the expected improvements in high-throughput sequencing technology in terms of read length and read accuracy, these issues may be of lesser importance in the future.

Acknowledgments

We wish to thank Dr. Junjie Qin from BGI-Shenzhen bioinformatics center for assisting us in accessing the BGI dataset, and for providing us with detailed explanations regarding the data generation process.

Author Contributions

Conceived and designed the experiments: YB EH. Performed the experiments: YB. Analyzed the data: YB. Wrote the paper: YB EH.

References

1. Rusch D, Halpern A, Sutton G, Heidelberg K, Williamson S, et al. (2007) The sorcerer II global ocean sampling expedition: northwest atlantic through eastern tropical pacific. PLoS Bio 5: e77.
2. Qin J, Li R, Raes J, Arumugam M, Burgdorf K, et al. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464: 59–65.
3. Huson D, Auch A, Qi J, Schuster S (2007) MEGAN analysis of metagenomic data. Genome Res 17: 377.
4. Hamady M, Walker J, Harris J, Gold N, Knight R (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. Nat Methods 5: 235–237.
5. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, et al. (2011) Enterotypes of the human gut microbiome. Nature 473: 174–180.
6. Karlin S, Mrazek J, Campbell A (1997) Compositional biases of bacterial genomes and evolutionary implications. J Bacteriol 179: 3899.
7. Takahashi M, Kryukov K, Saitou N (2009) Estimation of bacterial species phylogeny through oligonucleotide frequency distances. Genomics 93: 525–533.
8. Hofmann T (1999) Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval;15–19 August 1999; Berkeley, California, United States. SIGIR 99. Available: http://dl.acm.org/citation.cfm?id = 312649. ACM. pp 50–57.
9. Meinicke P, Aßhauer K, Lingner T (2011) Mixture models for analysis of the taxonomic composition of metagenomes. Bioinformatics 27: 1618.
10. Teeling H, Waldmann J, Lombardot T, Bauer M, Glöckner F (2004) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. BMC Bioinformatics 5: 163.
11. Chatterji S, Yamazaki I, Bai Z, Eisen J (2008) CompostBin: A DNA compositionbased algorithm for binning environmental shotgun reads. In: Proceedings of the 12th annual international conference on Research in computational molecular biology; 30 March–2 April, 2008; Singapore. RECOMB 2008. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi = 10.1.1.186.4259&rep = rep1&type = pdf. ISCB. pp 17–28.
12. Chan C, Hsu A, Tang S, Halgamuge S (2008) Using growing self-organising maps to improve the binning process in environmental whole-genome shotgun sequencing. J Biomed Biotechnol 2008: 513701.
13. Kislyuk A, Bhatnagar S, Dushoff J, Weitz J (2009) Unsupervised statistical clustering of environmental shotgun sequences. BMC Bioinformatics 10: 316.
14. Yang B, Peng Y, Leung H, Yiu S, Chen J, et al. (2010) Unsupervised binning of environmental genomic fragments based on an error robust selection of l-mers. BMC Bioinformatics 11: S5.
15. Leung H, Yiu S, Yang B, Peng Y, Wang Y, et al. (2011) A robust and accurate binning algorithm for metagenomic sequences with arbitrary species abundance ratio. Bioinformatics 27: 1489.
16. Wu Y, Ye Y (2010) A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. In: Proceedings of the 14th Annual International Conference on Research in Computational Molecular Biology; 12–15 August, 2010; Lisbon, Portugal. RECOMB 2010. Available: http://www.springerlink.com/content/2013001701185173. ISCB. pp 535–549.
17. Turnbaugh P, Hamady M, Yatsunenko T, Cantarel B, Duncan A, et al. (2008) A core gut microbiome in obese and lean twins. Nature 457: 480–484.
18. Jolliffe I (2002) Principal component analysis. In: Springer Series in Statistics, 2nd edition. 502 p.
19. Lee D, Seung H (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401: 788–791.
20. Jiang X, Weitz J, Dushoff J (2011) A non-negative matrix factorization framework for identifying modular patterns in metagenomic profile data. J Math Biol [Epub ahead of print].
21. Hofmann T (1999) Probabilistic latent semantic analysis. In: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence; 30 July–1 August 1999; Stockholm, Sweden. UAI 99. Available: http://uai.sis.pitt.edu/papers/99/p289-hofmann.pdf. AUAI. pp 289–296.
22. Hofmann T (2001) Unsupervised learning by probabilistic latent semantic analysis. Mach Learn 42: 177–196.
23. Cohn D, Chang H (2000) Learning to probabilistically identify authoritative documents. In: Proceedings of the Seventeenth International Conference on Machine Learning; 29 June–2 July, 2000; Stanford University, Stanford, CA, USA. ICML 2000. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi = 10.1.1.28.2691&rep = rep1&type = pdf. ACM. pp 167–174.
24. Cohn D, Hofmann T (2001) The missing link-a probabilistic model of document content and hypertext connectivity. In: Advances in neural information processing systems 13: proceedings of the 2000 conference; 27 November–2 December, 2000; Denver, Colorado, United States. NIPS 2000.Available: http://citeseer.ist.psu.edu/viewdoc/summary?doi = 10.1.1.33.6843. NIPS. pp 430–436.
25. Brants T, Chen F, Tsochantaridis I (2002) Topic-based document segmentation with probabilistic latent semantic analysis. In: Proceedings of the eleventh international conference on Information and knowledge management; 4–9 November, 2002; McLean, Virginia, United States. CIKM 02. Available: http://dl.acm.org/citation.cfm?id = 584829&bnc = 1. ACM. pp 211–218.
26. Sivic J, Russell B, Efros A, Zisserman A, Freeman W (2005) Discovering objects and their location in images. In: Proceedings of the tenth IEEE International Conference on Computer VisionComputer Vision. 17–20 October, 2005. Beijing, China. ICCV 2005. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp = &arnumber = 1541280. IEEE, volume 1. pp 370–377.
27. Richter D, Ott F, Auch A, Schmid R, Huson D (2008) Metasim - a sequencing simulator for genomics and metagenomics. PLoS One 3: e3373.
28. Kent W (2002) BLAT - the BLAST-like alignment tool. Genome Res 12: 656.
29. Chiang M, Mirkin B (2007) Experiments for the number of clusters in k-means. Progress in Artificial Intelligence 4874: 395–405.
30. Hartigan J (1975) Clustering Algorithms. New York: J. Wiley & Sons. 366 p.
31. Wendl M, Wilson R (2009) The theory of discovering rare variants via DNA sequencing. BMC Genomics 10: 485.