Published in final edited form as:

Stat Biosci. 2017 June ; 9(1): 13–27. doi:10.1007/s12561-016-9148-x.

A Model-Based Approach For Species Abundance Quantification Based On Shotgun Metagenomic Data

Eric Z. Chen¹, Frederic D. Bushman², and Hongzhe Li¹

¹Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA

²Department of Microbiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA

Abstract

The human microbiome, which includes the collective microbes residing in or on the human body, has a profound influence on the human health. DNA sequencing technology has made the large-scale human microbiome studies possible by using shotgun metagenomic sequencing. One important aspect of data analysis of such metagenomic data is to quantify the bacterial abundances based on the metagenomic sequencing data. Existing methods almost always quantify such abundances one sample at a time, which ignore certain systematic differences in read coverage along the genomes due to GC contents, copy number variation and the bacterial origin of replication. In order to account for such differences in read counts, we propose a multi-sample Poisson model to quantify microbial abundances based on read counts that are assigned to species-specific taxonomic markers. Our model takes into account the marker-specific effects when normalizing the sequencing count data in order to obtain more accurate quantification of the species abundances. Compared to currently available methods on simulated data and real data sets, our method has demonstrated an improved accuracy in bacterial abundance quantification, which leads to more biologically interesting results from downstream data analysis.

Keywords

Multi-Sample Poisson Model; Marker-Specific Effects; Microbiome; Read Coverage Variation

1 Introduction

The vast majority of the microorganisms on or in the human body inhabit in the gut. The collective genomes of these microbes are called microbiome. The gut microbiome plays important roles in human metabolism, nutrient intake and energy generation and is associated with many human diseases such as obesity, diabetes, and inflammatory bowel disease (IBD) (Cho and Blaser, 2012; Lewis et al., 2015). High-throughput sequencing technologies have been widely used to explore the microbial community in order to

Tel.: 215-573-5038, Fax: 215-573-4865, hongzhe@upenn.edu.

Software

We have implemented the method as an R package MSSQ, which is available on GitHub (<https://github.com/chvlyl/MSSQ>).

understand their roles in human health and diseases. One approach used in microbiome studies is based on 16S ribosomal RNA (rRNA) sequencing, which sequences the 16S rRNA gene to profile the bacterial community. The 16S rRNA gene sequence uniquely exists in prokaryotes and its high variability in microbial genomes can be exploited to identify different microbes. However, the 16S data are limited in discerning the bacteria at the species or strain level. Alternatively, shotgun sequencing of metagenomes, which sequences all genome sequences presented in the sample instead of just one marker gene, provides a more comprehensive approach to study human microbiome. This approach provides richer information about the microbial composition and gene functions. Both approaches are powerful and have been widely used in human microbiome studies (Turnbaugh et al., 2007; Qin et al., 2010).

One key problem in analysis of shotgun metagenomic sequencing data is to accurately and efficiently estimate the microbial abundances in the samples. Unlike the 16S rRNA sequencing approach that only involves one marker gene, the shotgun sequencing approach generates millions of reads from potentially all microbial genomes, which may share great similarity and also distribute unevenly in the sample. Therefore, analysis of shotgun metagenomic data is more complex and requires new statistical models and computational tools. To quantify microbial abundances, the majority of the current methods for metagenomic data analysis first align the sequencing reads to the known microbial genomes by sequence-alignment tools such as BLAST (Altschul et al., 1990) or Bowtie (Langmead et al., 2009). Since the microbial genomes presented in the sample share similarities, it is usually difficult to uniquely align back the sequencing reads that are generated from these similar genome regions. Aligning sequence reads of the whole metagenomes is also computationally demanding since each read has to be aligned to thousands of complete bacterial genomes.

To overcome the problem of read assignment ambiguity and to improve the computational efficiency, Segata et al. (2012) developed a computationally efficient algorithm, *MetaPhlAn*, which aligns reads only to the unique clade-specific marker genes that are identified from known microbial genomes and thus allows unambiguous taxonomic assignments (Segata et al., 2012). The relative abundances of the bacteria are then quantified only by the read counts that are aligned to these marker genes. However, for a given species, we observed that the reads are not evenly aligned to the markers and this marker-to-marker variability is consistent across all samples. Such observed marker-to-marker variability is possibly due to different GC contents, mappability and possible lateral gene transfers. *MetaPhlAn* ignores such marker effects and simply averages the normalized read counts of each marker to estimate the microbial abundances. Without considering such marker-specific effects, the abundance estimation can be biased.

Here, we propose a multi-sample species quantification (MSSQ) method based on Poisson model to account for the marker effects for robust estimation of microbial abundances based on shotgun sequencing data. In our proposed method, we develop a Poisson model with additional parameters to handle the marker effects, which are estimated by using multiple samples together. One innovation of our proposed method is to take into account the systematic bias of the taxonomic markers when estimating the microbial abundances. We

evaluate and compare the performances of proposed method with *MetaPhlAn*, a popular approach based on clade-specific marker genes for bacterial abundance quantification.

The paper is organized as follows. Section 2 introduces the multi-sample Poisson model to account for marker-to-marker variability of mapped read counts. Section 3 presents the simulation results to evaluate the MSSQ model and to compare it with *MetaPhlAn*. Section 4 applies the proposed method to a human gut microbiome data set to quantify the bacterial species associated with pediatric Crohn's disease (CD). Finally, we give a brief discussion of the methods and results in Section 5.

2 A Multi-Sample Poisson Model for Species Abundance Quantification

Our proposed method is based on the aligning of metagenomic read data into a set of clade-specific marker genes (Segata et al., 2012) or universal marker genes (Sunagawa et al., 2013), both have been shown to be effective in quantifying known microorganisms at species-level resolution using shotgun sequencing data. These methods simply average the normalized read counts, which are the read counts normalized by the marker lengths over all the markers to estimate the relative microbial abundances. However, we have observed that the number of aligned reads is not uniformly distributed across the marker genes within the same clade and sample (see Figure 1 for an illustration of such marker-specific effects). In order to account for such a marker-to-marker variability and estimate the microbial abundances, we propose a Poisson model for multiple sample analysis of read counts mapped to these marker genes.

2.1 A Multi-Sample Poisson Model to Account for Marker-Specific Effects

Consider a metagenomic study with N samples. After the sequencing reads are aligned to sets of clade-specific marker genes, the data can be summarized as a large table of counts as shown in Table 1, where X_{ijk} is the count data of sequencing reads for sample i ($i = 1, 2, \dots, n$), species j ($j = 1, 2, \dots, p$) and marker k ($k = 1, 2, \dots, m_j$). We model the count data for all species and all samples together and assume that the count X_{ijk} is generated from the following Poisson model,

$$X_{ijk} \sim \text{Poisson}(\theta_{ij} t_i \phi_{jk} l_{jk}), \quad (1)$$

where $\theta_{ij} > 0$ is the relative abundance for the j th species in the i th sample. In common practice, the bacterial abundance are usually transformed into relative abundances (i.e. the

bacterial abundance sum to 100% in one sample). We therefore impose that $\sum_{j=1}^p \theta_{ij} = 1$. This constraint also avoids the identifiability issue in the model. Here t_i is the total read counts for sample i that are mapped to the marker genes and l_{jk} is the length of the k th marker gene for j th species. Note that the species may have different number of markers. t_i and l_{jk} are known or can be calculated from the data directly. The parameters $\phi_{jk} > 0$ ($j = 1, \dots, p$ and $k = 1, \dots, m_j$) are used to model the marker-specific effects for the set of marker genes. The marker-specific effect can be due to different GC contents, mappability and

possible lateral gene transfers. Note that although the marker length parameters l_{jk} do not affect the estimates of θ_{ij} , including l_{jk} in the model makes the values of ϕ_{jk} comparable across markers and more interpretable. In fact, the variability of ϕ_{jk} across all m_j marker genes for a given species j can have an important biological meaning, as we demonstrate in our real data analysis.

In Model (1), for a given species j , each sample i has its own relative abundance θ_{ij} . The data X_{ijk} ($k = 1, \dots, m_j$) that we use to estimate θ_{ij} are assumed to follow a Poisson distribution. However, we allow each marker gene (k) to have its own effect, therefore, this actually accounts for possible overdispersion observed in the data. Our model uses data from multiple samples to estimate the marker effects ϕ_{jk} . When $\phi_{jk} = 1$, our model is a Poisson model without considering the marker effects, which is essentially the approach used by *MetaPhlAn*.

2.2 Model Fitting

We fit the model and estimate the parameter using the maximum likelihood estimation, where the likelihood function is

$$L = \prod_{i=1}^n \prod_{j=1}^p \prod_{k=1}^{m_j} \frac{e^{-\theta_{ij} t_i \phi_{jk} l_{jk}} (\theta_{ij} t_i \phi_{jk} l_{jk})^{x_{ijk}}}{x_{ijk}!},$$

and its logarithm is

$$-\log L = - \sum_{i=1}^n \sum_{j=1}^p \sum_{k=1}^{m_j} (-\theta_{ij} t_i \phi_{jk} l_{jk} + x_{ijk} \log(\theta_{ij} t_i \phi_{jk} l_{jk}) - \log(x_{ijk}!)). \quad (2)$$

Estimates θ and ϕ can be obtained iteratively. We first estimate all θ_{ij} by fixing ϕ_{jk} at the current values,

$$\hat{\theta}_{ij} = \frac{\sum_k^{m_j} x_{ijk}}{t_i \sum_k^{m_j} \phi_{jk} l_{jk}},$$

and then normalize these values so that $\sum_{j=1}^p \hat{\theta}_{ij} = 1$ for $i = 1, \dots, n$. This is equivalent to fitting a Poisson regression model based on the observed data $\{X_{ijk}, k = 1, \dots, m_j\}$. We then estimate all ϕ_{jk} by fixing θ_{ij} at the current values,

$$\hat{\phi}_{jk} = \frac{\sum_i^n x_{ijk}}{l_{jk} \sum_i^n \theta_{ij} t_i}.$$

This is equivalent to fitting a Poisson regression model based on the observed data $\{X_{ijk}, i = 1, \dots, n\}$.

We iteratively updates θ_{ij} and ϕ_{jk} until convergence. Since the closed-form updating formula are available, the algorithm is very efficient. The objective function (2) is biconvex, but not convex. As in any non-convex optimization problem, there is no guarantee that a global minimum can be achieved. From our simulations, if we start with *MetaPhlAn* estimates with all ϕ_{jk} being set to 1, we always obtain the correct estimates.

3 Simulation Studies

We first performed a simulation study to examine the model fitting, where we simulated different sample size $n = (20, 50, 100)$ and different number of markers $m = (20, 50, 100)$ per species for a total of $p = 50$ species. In order to mimic the skewed distribution of species abundance in a sample, we simulated θ_{ij} from a log-normal distribution with $\mu = 0$ and $sd =$

1.5. The simulated θ_{ij} were then normalized so that $\sum_{j=1}^p \theta_{ij} = 1$. The scaled marker length l_{jk} and total read counts t_i was simulated uniformly from $[0.1, 10]$ and $[50, 500]$, respectively. We simulated ϕ_{jk} from a log-normal distribution with $\mu = 0$ and $sd = 1.5$. We fitted the MSSQ model on the simulated data and then calculate the average relative difference (AVGRD) for $\hat{\theta}$ and $\hat{\phi}$, which is defined as

$$\frac{1}{n} \sum_{i=1}^n \frac{|\theta_{ij} - \hat{\theta}_{ij}|}{\hat{\theta}_{ij}}, \frac{1}{m} \sum_{k=1}^m \frac{|\phi_{jk} - \hat{\phi}_{jk}|}{\hat{\phi}_{jk}}.$$

The simulations were repeated 1,000 times for each combination of sample size and marker size. The mean and standard error of the AVGRD over 1000 repeated simulations were then calculated and reported.

Table 2 shows the simulation results. When the number of markers per species is fixed, increasing sample size improves the estimation accuracy of both θ and ϕ . Similarly, when the sample size is fixed, more markers can improve the estimation accuracy of both θ and ϕ .

We next studied the effects of allowing the marker-specific effects in the Poisson on species abundance quantification, where we simulated $n = 50$ samples and each sample had $p = 50$ species. The true abundance θ_{ij} , the scaled marker length l_{jk} and total read counts t_i were simulated in the same way as previously. For each species, the number of markers m_j was uniformly simulated from $[10, 50]$. To mimic different degrees of marker effects, we simulated ϕ_{jk} from a log-normal distribution with $\mu = 0$ and $sd = (0, 0.5, 1, 1.5)$ for four different simulation settings, respectively. The simulation setting with log-normal ($\mu = 0$, $sd = 0$) indicates no marker effect in the data and the other three settings indicate increasing marker effects.

We observed that ignoring the marker-specific effects led to large biases in the estimates of the abundance θ_{ij} as shown in Figure 2 (full-scale in upper panel and zoomed-in to 0–5% scale in lower panel). When the marker effect increases, the estimated abundance deviated

more from the true abundance, where the mean of AVGRD were 0.07, 0.15, 0.28, 0.52. In contrast, estimates of the relative abundances from MSSQ showed little biases when there were strong marker-specific effects. The mean of AVGRD were 0.07, 0.09, 0.09, 0.10, respectively.

4 Application to the human gut microbiome data

4.1 Data description and processing

We applied MSSQ to a data set of shotgun metagenomic study comparing gut microbiome between healthy children and children with Crohn's disease (Lewis et al., 2015; Lee et al., 2015). A total of 90 children with Crohn's disease and 26 healthy controls were recruited to the study and provided fecal samples for shotgun metagenomic sequencing. DNAs were prepared from whole stool and were sequenced using the Illumina HiSeq paired-end method, resulting in an average number of reads per sample of 11×10^6 with an average length 100 bases on each end of the paired-end reads. We were interested in quantifying the bacterial abundances and identifying the bacteria species that show differential abundance between the healthy control samples and the Crohn's disease samples.

We first aligned the sequencing reads to the *MetaPhlAn* clade specific markers for each of the bacterial species. In order to compare our method with *MetaPhlAn*, we applied the same filtering criteria used by *MetaPhlAn*. Specifically, for each species, we filtered out the upper 10% percentile of the most abundant markers and lower 10% percentile least abundant markers. We replaced the read counts of those filtered markers with NAs and our model ignores the NAs in the model fitting procedure. We ran *MetaPhlAn* on the same data set and compare the results with those from MSSQ.

4.2 Abundance estimation with MSSQ

After the filtering steps, we applied MSSQ to quantify the species abundance from healthy children and those with Crohn's disease separately. MSSQ identified a total of 138 species presented in the healthy control samples and 216 in the disease samples. *MetaPhlAn* identified the same number of species in control and disease samples. Figure 3 shows the estimated marker effect $\hat{\phi}$ for each species in control samples and disease samples. The dispersion of $\hat{\phi}$ values as measured by the median absolute deviation (MAD) indicates that the marker effects are indeed present in the real data. The median dispersion of $\hat{\phi}$ is 2.77 for species in control samples and 2.68 for species in disease samples.

To further demonstrate the relevance of the marker-specific effects ϕ_{jk} , Figure 4 shows a comparison of the distributions of the estimated marker-specific effects in the control samples and disease samples for 23 species reported in Korem et al. (2015) whose peak-to-trough ratios (PTRs) are associated with several diseases and metabolic disorders, where PTR was used to reflect the growth rate of the bacterial population. Korem et al. (2015) found strong associations between the PTRs of two species, *Escherichia coli* and *Bifidobacterium longum*, and Crohn's disease. Interestingly, our analysis also showed very significant differences in the estimated marker-specific effects for these two species between healthy control and Crohn's disease samples, indicating that such marker-specific effects

may reflect the bacterial growth rate. Our analysis also indicated that *Ruminococcus torques* showed a large difference in the estimated marker-specific effects. It is also interesting to observe that *Klebsiella pneumoniae* showed an opposite difference, having a larger estimated marker-specific effect in the healthy controls than the cases.

We then compared the abundance estimation by MSSQ and *MetaPhlAn* (Figure 5). When the marker-specific effects were excluded from the model, i.e., setting all $\phi_{jk}=1$, as expected, MSSQ resulted in the same abundance estimate as the *MetaPhlAn* for both healthy control and disease samples. When the marker-specific effects were estimated, MSSQ shows different abundance estimation from *MetaPhlAn*. Overall, we observed that the estimates from both methods were comparable with a Pearson correlation of 0.92 (See Figure 5 upper panel). However, for low abundant species, a large discrepancies of the estimates were observed (Figure 5 lower panel). Due to low read counts to these species, a large variability in these estimates was expected. The overall pattern was similar to our simulation results in Figure 2. We also observed that the differences were greater in disease samples than in the healthy control samples.

4.3 Clustering and differential abundance analysis

We next evaluated whether the abundance quantification by MSSQ can improve the downstream analysis such as clustering and differential abundance analysis. Before we performed the downstream analysis, we first filtered out the rare species, which is a common pre-processing procedure often used in the literature (Kostic et al., 2015; Romero et al., 2014; Stein et al., 2013). Particularly, the species with an estimated abundance $< 1\%$ in all the control and disease samples in both MSSQ and *MetaPhlAn* estimation were filtered out. After filtering, 136 species were left for following analysis.

Figure 6 shows the heatmap of the estimated abundances from MSSQ and *MetaPhlAn*, where the samples were clustered with Bray-Curtis distance and the species were clustered based on Pearson's correlations. We observed a clear separation of control and Crohn's disease gut microbiome samples based on *MetaPhlAn* and MSSQ estimation of the species abundances. However, the clustering results based on the MSSQ estimation show a clearer pattern. For example, the species within the clusters estimated from MSSQ seem to have greater similarity in their abundances across all samples than those estimated from *MetaPhlAn*.

As a comparison, we also tested for differential abundance between control and disease samples for each species using the Wilcoxon rank-sum test. Using these estimated abundance from *MetaPhlAn*, Wilcoxon rank-sum test identified 39 differentially abundant species with FDR adjusted p -value < 0.05 . Based on the estimated abundance from MSSQ, 41 species were identified to be differentially abundant between control and disease samples, including all 39 differentially abundant species identified based on the *MetaPhlAn* estimation. Among the species that are more abundant in Crohn's patients, *E. coli*, *Bacteroides*, *Enterococcus*, *Klebsiella*, and *Ruminococcus gnavus* have been reported in literature (Sartor, 2006; Liu et al., 1995). Among the protective species, *Bifidobacterium*, *Roseburia* and *Eubacterium* were also reported in previous studies (Manichanh et al., 2012).

Manichanh et al. (2012) detected *Lactobacillus* as a protective species that was more abundant in normal gut.

The two species that were only identified when MSSQ was applied include *Coprococcus comes*, which belongs to genera *Coprococcus*, and *Clostridium symbiosum*. Both species were reported to be associated with dysbiosis, CD (Gevers et al., 2014) and UC (Du et al., 2015). *Clostridium symbiosum* has been reported to protect the gut mucosa by producing butyrate (Van den Abbeele et al., 2013).

5 Discussion

We have proposed a multi-sample Poisson model to quantify the bacterial abundances based on the shotgun metagenomic data. Our method uses the count data of reads aligned to clade-specific marker genes to quantify the species abundances, taking into account the marker effects that are observed in the shotgun sequencing data. In our analysis of real data, we used the marker genes sets used in *MetaPhlAn* package. Alternatively, one can also use the 40 universal marker genes defined by Sunagawa et al. (2013). It would be interesting to compare whether different marker genes lead to similar species quantification. Simulation results indicate that the proposed model-based approach can lead to better quantification of species abundances.

We further demonstrated the proposed methods in an analysis of pediatric gut microbiome to identify the species that are associated with Crohn's disease. The discrepancy of the differential abundant species identified by MSSQ and *MetaPhlAn* is due to whether the marker-specific effects are accounted in the estimation. Our simulations have clearly demonstrated that ignoring such marker-specific effects can lead to biased quantification of the species abundances. The marker-specific effects could originate from different sources. For example, it is well known that the sequencing reads are not uniformly distributed among genomic regions and are correlated with GC contents of the genomic regions (Ross et al., 2013). Furthermore, the copy number variation of the genes can also contribute to the under- or over- presentation of certain gene regions. Sequencing error and alignment error are other potential sources for non-uniform distribution of the reads and may cause the large marker-to-marker variability we observed in the data.

In our analysis, we fitted the multi-sample Poisson model for all the species and all the samples simultaneously. Since we do not expect to observe all the species in all the samples, some regularization on the relative abundance parameter θ_{ij} may help to improve the abundance quantification. Finally, it is interesting to establish the asymptotic distribution of the resulting estimates of the relative abundances when both the sample size n and the numbers of marker genes m_j go to infinity. The difficulty is that the variances of $\hat{\theta}_{ij}$ and $\hat{\phi}_{jk}$ depend on m_j and n . However, as $m_j \rightarrow \infty$ and $n \rightarrow \infty$, the number of the parameters in the model diverges and also increases to ∞ , which makes the asymptotic analysis difficult. In addition, $\hat{\theta}_{ij}$ and $\hat{\phi}_{jk}$ are estimated iteratively and do not have explicit expressions.

Acknowledgments

Supported by NIH grants CA127334 and GM097505.

References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *Journal of molecular biology*. 1990; 215:403–410. [PubMed: 2231712]
- Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. *Nature Reviews Genetics*. 2012; 13:260–270.
- Du Z, Hudcovic T, Mrazek J, Kozakova H, Srutkova D, Schwarzer M, Tlaskalova-Hogenova H, Kostovcik M, Kverka M. Development of gut inflammation in mice colonized With mucosa-associated bacteria From patients With ulcerative colitis. *Gut pathogens*. 2015; 7:1. [PubMed: 25653718]
- Gevers D, Kugathasan S, Denson LA, Vázquez-Baeza Y, Van Treuren W, Ren B, Schwager E, Knights D, Song SJ, Yassour M, et al. The treatment-naïve microbiome in new-onset Crohns disease. *Cell host & microbe*. 2014; 15:382–392. [PubMed: 24629344]
- Korem T, Zeevi D, Suez J, Weinberger A, Avnit-Sagi T, Pompan-Lotan M, Matot E, Jona G, Harmelin A, Cohen N, Sirota-Madi A, Thaïss CA, Pevsner-Fischer M, Sorek R, Xavier R, Elinav E, Segal E. Growth dynamics of gut microbiota in health and disease inferred From single metagenomic samples. *Science*. 2015; 349:1101–1106. [PubMed: 26229116]
- Kostic AD, Gevers D, Siljander H, Vatanen T, Hyötyläinen T, Hämäläinen AM, Peet A, Tillmann V, Pöhö P, Mattila I, Lähdesmäki H, Franzosa EA, Vaarala O, de Goffau M, Harmsen H, Ilonen J, Virtanen SM, Clish CB, Oreši M, Huttenhower C, Knip M, Xavier RJ. The Dynamics of the Human Infant Gut Microbiome in Development and in Progression toward Type 1 Diabetes. *Cell host & microbe*. 2015; 17:260–273. [PubMed: 25662751]
- Langmead B, Trapnell C, Pop M, Salzberg SL, et al. Ultra-fast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009; 10:R25. [PubMed: 19261174]
- Lee D, Baldassano R, Otley A, Albenberg L, Griffiths A, Compher C, Chen E, Li H, Gilroy E, Nessel L, et al. Comparative Effectiveness of Nutritional and Biological Therapy in North American Children With Active Crohn's Disease. *Inflammatory bowel diseases*. 2015; 21:1786–1793. [PubMed: 25970545]
- Lewis JD, Chen EZ, Baldassano RN, Otley AR, Griffiths AM, Lee D, Bittinger K, Bailey A, Friedman ES, Hoffmann C, et al. Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohns Disease. *Cell Host & Microbe*. 2015; 18:489–500. [PubMed: 26468751]
- Liu Y, van Kruiningen H, West A, Cartun R, Cortot A, Colombel J. Immunocytochemical evidence of *Listeria*, *Escherichia coli*, and *Streptococcus* antigens in Crohn's disease. *Gastroenterology*. 1995; 108(5):1396–1404. [PubMed: 7729631]
- Manichanh C, Borruel N, Casellas F, Guarner F. The gut microbiota in IBD. *Nature Reviews Gastroenterology and Hepatology*. 2012; 9:599–608. [PubMed: 22907164]
- Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*. 2010; 464:59–65. [PubMed: 20203603]
- Romero R, Hassan SS, Gajer P, Tarca AL, Fadrosch DW, Nikita L, Galuppi M, Lamont RF, Chaemsaitong P, Miranda J, Chaiworapongsa T, Ravel J. The composition and stability of the vaginal microbiota of normal pregnant women is different From that of non-pregnant women. *Microbiome*. 2014; 2:4. [PubMed: 24484853]
- Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB. Characterizing and measuring bias in sequence data. *Genome Biol*. 2013; 14:R51. [PubMed: 23718773]
- Sartor R. Mechanisms of Disease: pathogenesis of Crohn's disease and ulcerative colitis. *Nature Clinical Practice Gastroenterology & Hepatology*. 2006; 3:390–407.
- Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods*. 2012; 9:811–814. [PubMed: 22688413]

- Stein RR, Bucci V, Toussaint NC, Buffie CG, Räscher G, Pamer EG, Sander C, Xavier JB. Ecological Modeling From Time-Series Inference: Insight into Dynamics and Stability of Intestinal Microbiota. *PLoS Computational Biology*. 2013; 9:e1003388. [PubMed: 24348232]
- Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR, Coelho LP, Arumugam M, Tap J, Nielsen HB, et al. Metagenomic species profiling using universal phylogenetic marker genes. *Nature methods*. 2013; 10:1196–1199. [PubMed: 24141494]
- Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. *Nature*. 2007; 449:804–810. [PubMed: 17943116]
- Van den Abbeele P, Belzer C, Goossens M, Kleerebezem M, De Vos WM, Thas O, De Weirde R, Kerckhof F-M, Van de Wiele T. Butyrate-producing *Clostridium* cluster XIVa species specifically colonize mucins in an in vitro gut model. *The ISME journal*. 2013; 7:949–961. [PubMed: 23235287]

Fig. 1. Marker effects in the shotgun metagenomic data for four bacterial species. The raw sequencing reads were first aligned to the taxonomic markers and then normalized as the number of aligned reads for each marker divided by marker length and total number of aligned reads for each sample. The normalized data was then clustered and showed in heatmap. The rows are samples and columns are taxonomic markers from *MetaPhlAn*.

Fig. 2.

Comparison of relative abundance estimation by MSSQ with and without estimating marker-specific effects for four scenarios with increasing marker-specific effects (from left to right), where the plots in the first column correspond to the model assuming no marker-specific effects ($\phi_{jk} = 1$ for all $j = 1, \dots, p$ and $k = 1, \dots, m_j$). A total of 50 samples, each with 50 species were simulated. Upper panel: full scale with all species; lower panel: zoom-in 0–5% scale with only the species with the relative abundances 1–5%.

Fig. 3. Distribution of estimated marker-specific effects for each species in the healthy control samples and disease samples. Each dot represents the estimated ϕ with log transformation and each boxplot represents a species. A total of 43 species are divided into top and bottom panels for visualization.

Fig. 4. Comparison of the distributions (boxplots) of estimated marker-specific effects in the healthy control samples and disease samples for 23 species reported in Korem et al. (2015) whose peak-to-trough ratios are associated with several diseases and metabolic disorders.

Fig. 5. Comparison of estimated relative abundances from MSSQ and *MetaPhlAn*. Upper panel: all species; Lower panel: species with abundances < 5%.

Fig. 6. Heatmap of the estimated relative abundances from MSSQ and *MetaPhlAn* based on analysis of metagenomic data comparing 26 control samples and 90 samples with Crohn's disease, where columns represent samples and rows represent bacterial species, and the clinical meta-data are presented in the top bars. See Lewis et al. (2015) and Lee et al. (2015) for details of the clinical meta-data.

Summary of count data after the sequencing reads from n metagenomic samples are aligned to a set of m_j clade-specific marker genes for the i th species for a total of p species, where X_{ijk} is the number of reads that are assigned to the k th marker gene of the j th species in the i th sample.

Table 1

	species 1				...	species j				...	species p			
	1	2	...	m_1		1	2	...	m_j		1	2	...	m_p
l_{i1}	i_{11}	i_{12}	...	l_{1m1}		l_{j1}	l_{j2}	...	l_{jmj}		l_{p1}	l_{p2}	...	l_{pm_p}
1														
2														
$\vdots$														
i						$X_{ijk} \Rightarrow \theta_{ij}$								
$\vdots$														
n														

$$\begin{aligned} i &= 1, 2, \dots, n \\ j &= 1, 2, \dots, p \\ k &= 1, 2, \dots, m_j \end{aligned}$$

Table 2

Parameter estimation for data simulated from the multi-sample Poisson model with different sample sizes n and different marker sizes m with number of species $p = 20$. For each parameter, the mean and standard error (SE) of average relative difference (AVGRD) are calculated based on 1000 replications. For each n and m , the first row presents the estimates of θ , the second row presents the estimates of ϕ .

AVGRD	$m = 20$		$m = 50$		$m = 100$		
	mean	SE	mean	SE	mean	SE	
$n=20$	θ	0.19	0.05	0.14	0.04	0.11	0.03
	ϕ	0.22	0.07	0.16	0.05	0.13	0.03
$n=50$	θ	0.09	0.02	0.07	0.01	0.06	0.01
	ϕ	0.08	0.02	0.07	0.02	0.07	0.01
$n=100$	θ	0.06	0.01	0.05	0.01	0.04	0.01
	ϕ	0.04	0.01	0.04	0.01	0.04	0.01