BST281: Genomic Data Manipulation, Spring 2018

Final project report

Application of Quantitative Microbiome Profiling (QMP) on Longitudinal Crohn's Disease Data from NIH Human Microbiome Project

Eunice Yeh, Tian Zhang, Marina Cheng, Anthony Lamattina; data courtesy of Casey DuLong, Tiffany Poon, Jason Lloyd-Price, Curtis Huttenhower, the Inflammatory Bowel Disease Multi'omics Database Team; and special thanks to Eric Franzosa for guidance

The gut microbiome has been a recent focus in the study of human health as it composes the largest and most influential collection of microorganisms in the human body. Its structural and functional relationships with various diseases or health states are of particular interest in improving diagnosis (i.e. bacterial markers) or developing novel therapeutic interventions (i.e. probiotics) for human diseases. However, more efforts are needed in understanding how the vast compositional variations in the gut microbes alter not only between individuals but also within an individual over time, along with various host phenotypes (i.e. the immune system) and host health status. Currently, the most common approach to measuring gut microbial community variation has been relative microbiome profiling (RMP), which quantifies the abundances of microbial taxa as fractions of the sample sequence library depending on the sequencing analysis. A major limitation of this method is that differences or changes in relative abundances cannot be accurately compared across samples with varying total microbial loads, which often occurs between samples from different individuals, especially of different health statuses, but can also even occur between samples collected from the same individual but at different time points. Thus, quantitative microbiome profiling (QMP) for absolute taxa abundances has been proposed recently by a few microbiome researchers[1,2,3,4]. Specifically inspired by the QMP approach presented by Vandeputte et al, for this project, we applied our own variant of this method on 22 longitudinal stool samples collected from a patient with Crohn's Disease (CD) and another 22 from a patient without any inflammatory bowel disease (IBD), as the healthy control for disease state comparisons, over the course of roughly 12 months as part of the NIH Human Microbiome Project.

The relative taxonomic profiles of the 44 stool samples were calculated as fractions, processed from a metagenomic analysis, and stored in the IBD Multi'omics Database (IBDMDB); data and protocol for the sequencing analysis is publicly accessible online at ibdmdb.org. To contrast this RMP method, our QMP approach is to quantify taxa abundances (focusing at the species-level) by multiplying these relative taxonomic profiles by the copy-number-corrected bacterial 16S DNA concentrations in the microbial sample. The raw dilution-adjusted quantifications of the bacterial DNA concentration (ng/uL) for the 44 stool samples were enumerated through qPCR (assays were performed in duplicates, following standard protocols) and provided directly to us by Curtis Huttenhower's lab. To adjust for multiple copies of the 16S rRNA gene, the average copy number variation in most bacteria were extracted from the Ribosomal RNA Database (rrnDB)[5]; for bacteria found in our samples without copy number information, we used the average of all the copy number means from rrnDB (which was about 3.2). Since the relative species profiles were calculated from metagenomic reads (as opposed to PCR amplification of a single gene), we needed to normalize the bacterial 16S DNA concentrations to one copy number by dividing each raw concentration from the qPCR by the total relative bacterial abundance that has accounted for the copy number variation across the existing bacteria within each sample (i.e. QMP = RMP × [qPCR concentration ÷ ∑(RMP × copy number)]). Thus, the resulting absolute abundance of each species in a sample is calculated by quantifying their relative profiled abundance to the concentration (ng/uL) of one bacterial 16S DNA copy in the sample.

Before applying our QMP approach, we must first verify whether the RMP method used on these 44 samples bypassed its major limitation mentioned above. If we can show that the metagenomic sequencing reads surely reflected the total microbial concentrations across all the samples, then QMP becomes irrelevant for the current data at hand. So we counted the number of sequenced reads from the individual raw FASTq files of

the 44 samples from IBDMDB. Separately, we found insufficient evidence for a statistically significant association between the number of sequenced reads and the microbial concentrations across samples from the CD patient (N=22, Spearman's rho=0.13, p=0.57) and the healthy control (N=22, Spearman's rho=-0.23, p=0.21) at the nominal alpha level of 0.05. Thus, we have justified the need to apply QMP on these 44 samples, hence moving forward with our analysis to compare QMP results against those obtained using RMP.

Our primary analysis is to compare the relative to quantitative species abundances for each of the two patients, first separately (**fig. 1**), then compared against each other but focusing on the particular species of interest (**fig. 2**). We would expect E. coli to be more abundant in the CD patient since it is the well-known culprit for IBD, explaining why it is one of the top 15 abundant species only for the CD patient and not for the healthy control. Similarly, most Prevotella bacteria are known to be good, often abundantly seen in a healthy human gut microbiome. Since our healthy control exhibited high abundances of P. copri in particular across most time points, we will focus on this species of Prevotella for diagnosis comparisons.

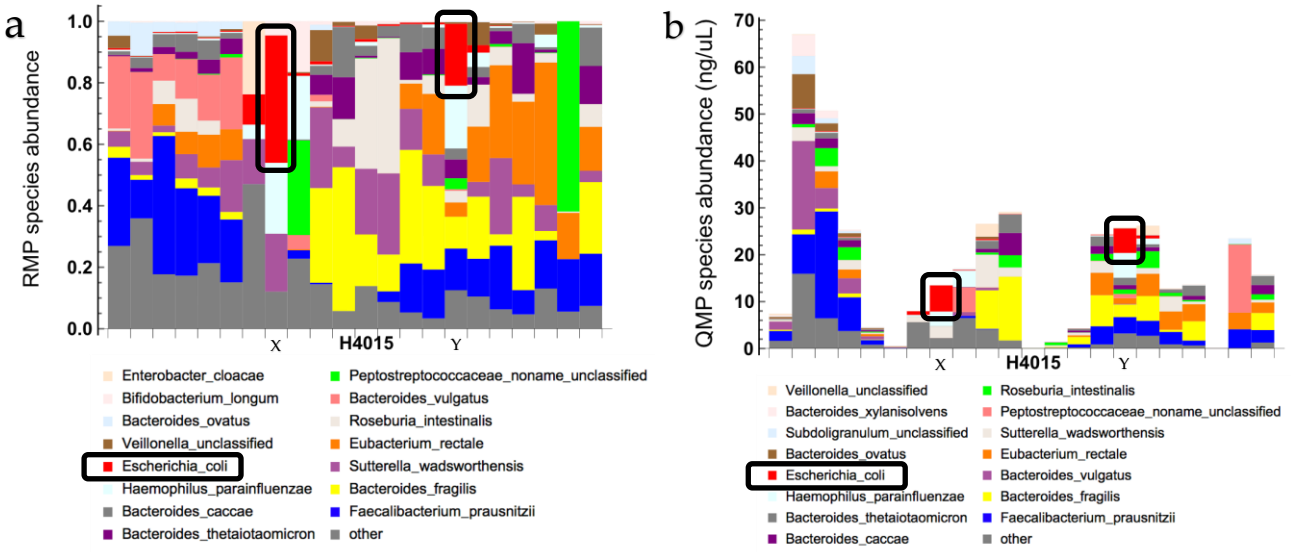


Figure 1: (a) Relative versus (b) quantitative microbiome profiling. Vertical bars represent species-level microbial composition of stool samples from the Crohn's Disease patient (H4015), ordered by sampling time. The top 15 most abundant species within this patient separately in RMP and QMP are colored, all others are grayed out and grouped into "other". The relative abundances of species in (a) are overestimated when the total microbial concentration in the sample is low as shown in (b); i.e. RMP of Escherichia coli (E. coli) at time X is distinctively greater than that at time Y, whereas the QMP of E. coli at the time X appears to be quite similar in absolute concentration to that at time Y; this is due to the total microbial concentration in the sample being lower at time X compared to time Y. Particularly in (b), a general decreasing trend in total microbial concentrations across samples (i.e. local maximums/peaks are lower over time) raises the question of whether the disease could be depleting the overall microbiota abundance itself, which would impact longitudinal comparisons among disease-associated health states. As one anecdotal contrast, the total microbial concentration consistently peaks between 25-35 ng/uL while fluctuating wildly nonetheless across the healthy control samples over time (see Supplementary Figure 1).

Our results demonstrated that even within an individual, taxa abundances between two samples collected at different time points can only be accurately compared on an absolute quantitative scale, as opposed to relative, when the total microbial load, or concentration, varies between the samples. Furthermore, differences in taxa abundances between two individuals of different disease states can be overestimated when measured by RMP, as compared to QMP. In fact, the average (log10) abundance of P. copri is significantly higher across samples from the healthy control than the CD patient when measured with RMP (p=0.024); however, this signal is lost using QMP (p=0.053). This result may be a consequence of the underlying limitation for RMP, which ignores the varying total microbial concentrations between the samples being compared.

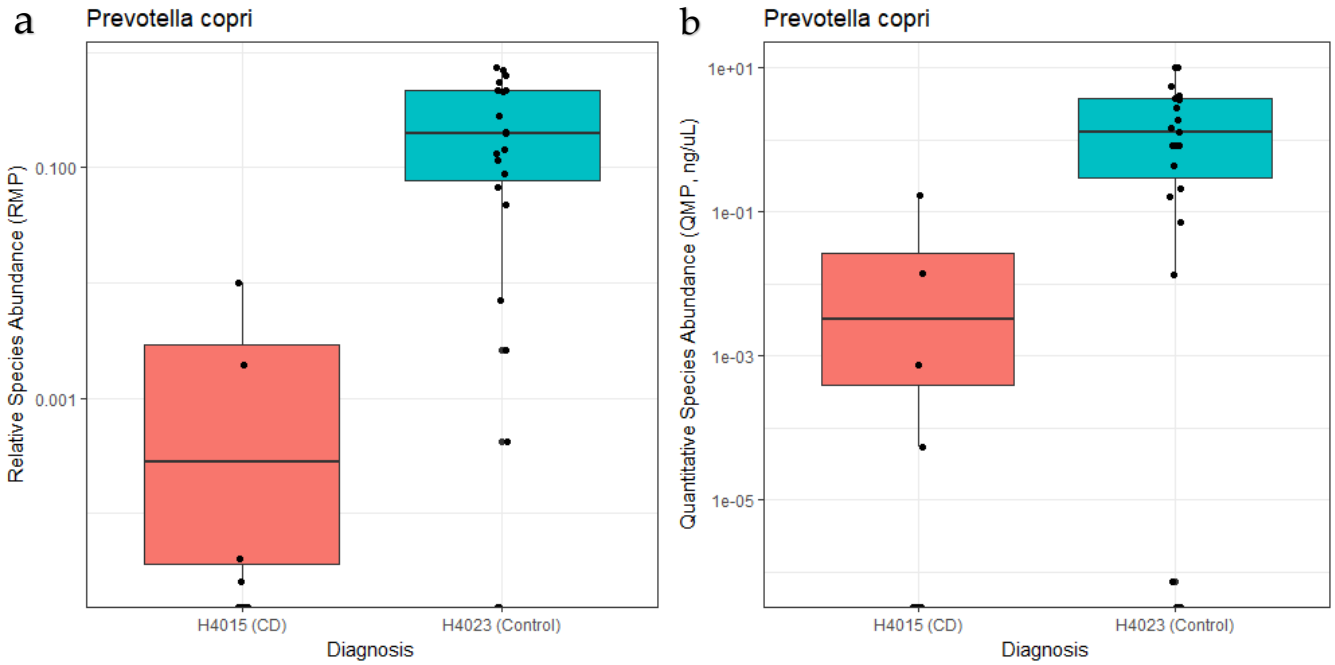


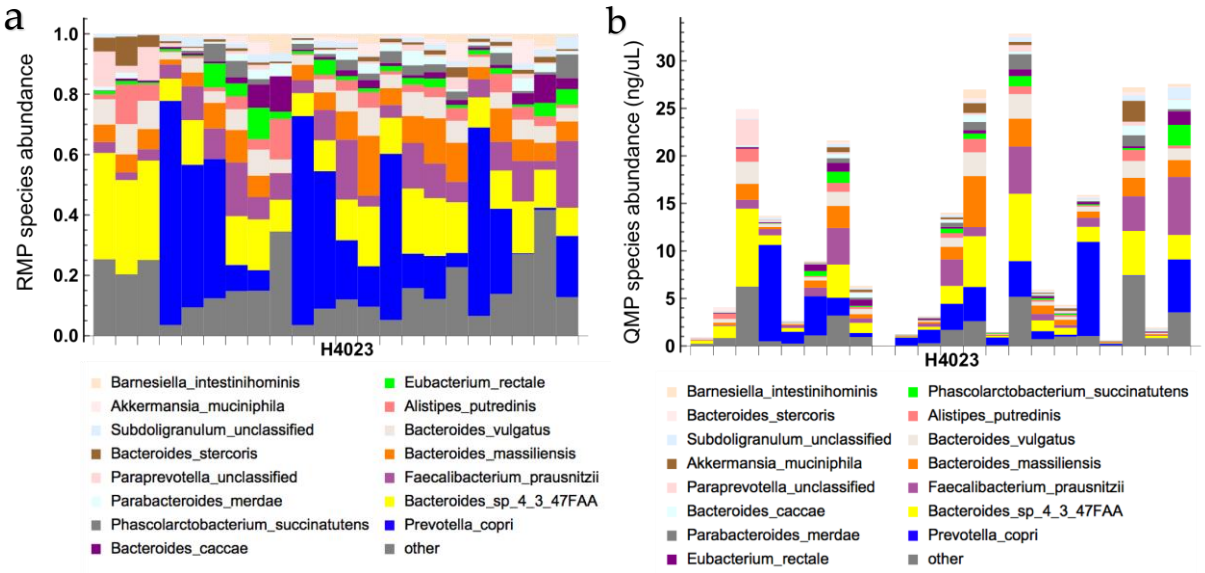
Figure 2: Distribution of (a) relative versus (b) quantitative P. copri abundances between the CD patient and healthy control. Combing longitudinal collection of the stool samples for each patient, the difference in P. copri abundances between the CD patient and the healthy control is more apparent when measured in (a) RMP compared to (b) QMP (both y-axes are log10-transformed). This difference is formally tested on the average of log10-transformed abundances between the two patients and is statistically significant when measured by RMP (p=0.024), while no longer found but is still marginally different by QMP (p=0.053); Welch's two-sample t-test at α=0.05. Similar comparisons in E. coli abundances is shown in Supplementary Figure 2.

As a supporting analysis, we showed that the significant differences seen in both relative abundances of P. copri and E. coli between our two selected patients were also represented in the rest of the study sample population (so comparing between all available samples from the CD patients and the non-IBD patients as healthy controls in IBDMDB). Based on the two-way ANOVA comparisons of the log10-transformed relative abundances between all CD patients and healthy controls, adjusted for repeated longitudinal samples within each patient, the difference is statistically significant both for P. copri ($p < 2\times10^{-16}$) and E. coli ($p = 1.7\times10^{-9}$) at the nominal alpha level of 0.05. Provided the resources to run more qPCR or even flow cytometric load assessment, an important future direction for this project would be to test whether this prominent effect of CD diagnosis on P. copri abundances remains true when calculated using QMP, and if so, would the size of the effect be more or less?

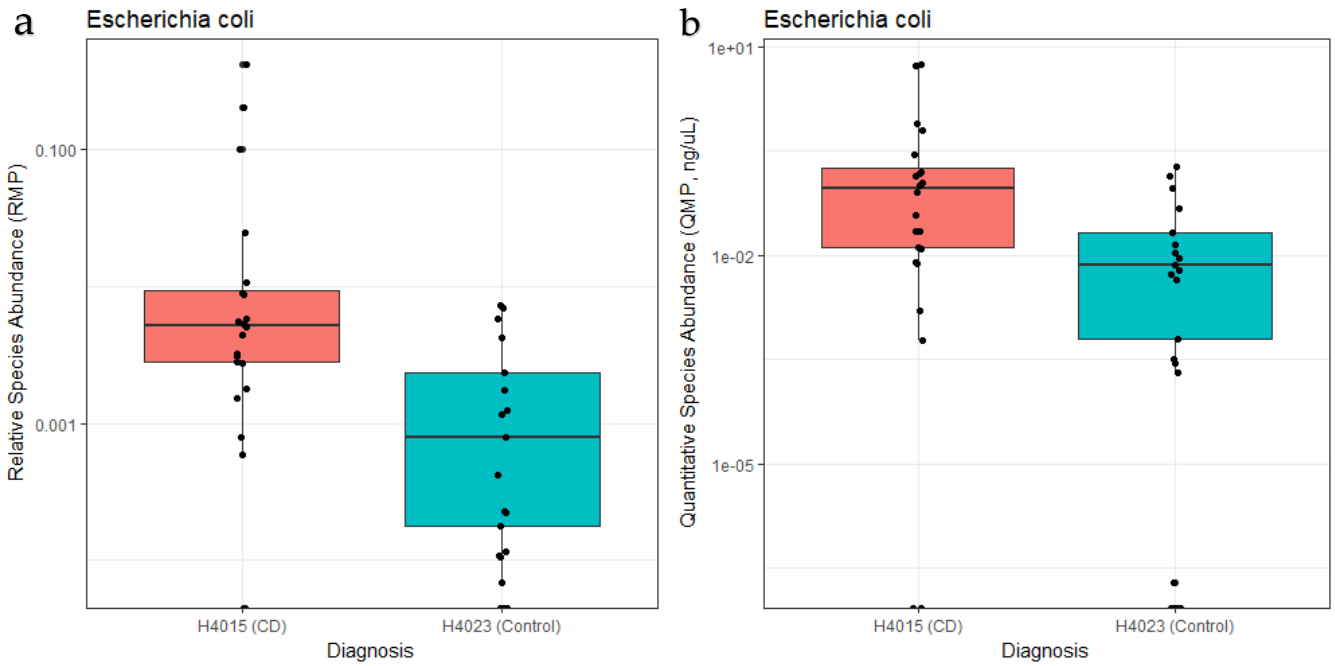
Further supplementary data summaries and findings can be found on https://github.com/euniceyeh/QMP-Project, which also includes the open source implementation of our QMP method among other programs on data derivation and analysis.

References

1. Satinsky, B. M., Gi-ord, S. M., Crump, B. C. & Moran, M. A. Use of internal standards for quantitative metatranscriptome and metagenome analysis. *Methods Enzymol.* 531, 237–250 (2013).
2. Props, R. et al. Absolute quanti cation of microbial taxon abundances. *ISME J.* 11, 584–587 (2017).
3. Stämmler, F. et al. Adjusting microbiome pro les for di-erences in microbial load by spike-in bacteria. Microbiome 4, 28 (2016).
4. Vandeputte, D. et al. Quantitative microbiome profiling links gut community variation to microbial load. *Nature.* **551,** 507–511 (23 November 2017).
5. Stoddard S.F, Smith B.J., Hein R., Roller B.R.K. and Schmidt T.M. (2015) rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. *Nucleic Acids Research* (2014).



Supplementary Figure 1: (a) Relative versus (b) quantitative microbiome profiling in CD patient over time. Vertical bars represent species-level microbial composition of stool samples from the healthy non-IBD patient (H4023), ordered by sampling time. The top 15 most abundant species within this patient separately in RMP and QMP are colored, all others are grayed out and grouped into "other". Supporting our main figure (fig. 1), species abundances in (a) RMP are also considered exaggerated when compared to abundances measured in (b) QMP across the healthy control samples. In particular, the visbly large fluctuation of Prevotella copri (P. copri) abundances in RMP is no longer apparent when measured in absolute concentrations (QMP); most likely due to the complementary fluctuation in total microbial concentrations across the samples.



Supplementary Figure 2: Distribution of (a) relative versus (b) quantitative E. coli abundances between the CD patient and healthy control. Combing longitudinal collection of the stool samples for each patient, the difference in E. coli abundances between the CD patient and the healthy control is more apparent when measured in (a) RMP compared to (b) QMP (both y-axes are log10-transformed). This difference is formally tested on the average of log10-transformed abundances between the two patients and reached statistical significance for both methods: RMP ($p<0.001$) vs. QMP ($p=0.006$); Welch's two-sample t-test at $\alpha=0.05$.