16S rRNA gene-based
Microbiome Taxonomic Profiling (MTP)

16S copy number correction

By Jon Jongsik Chun

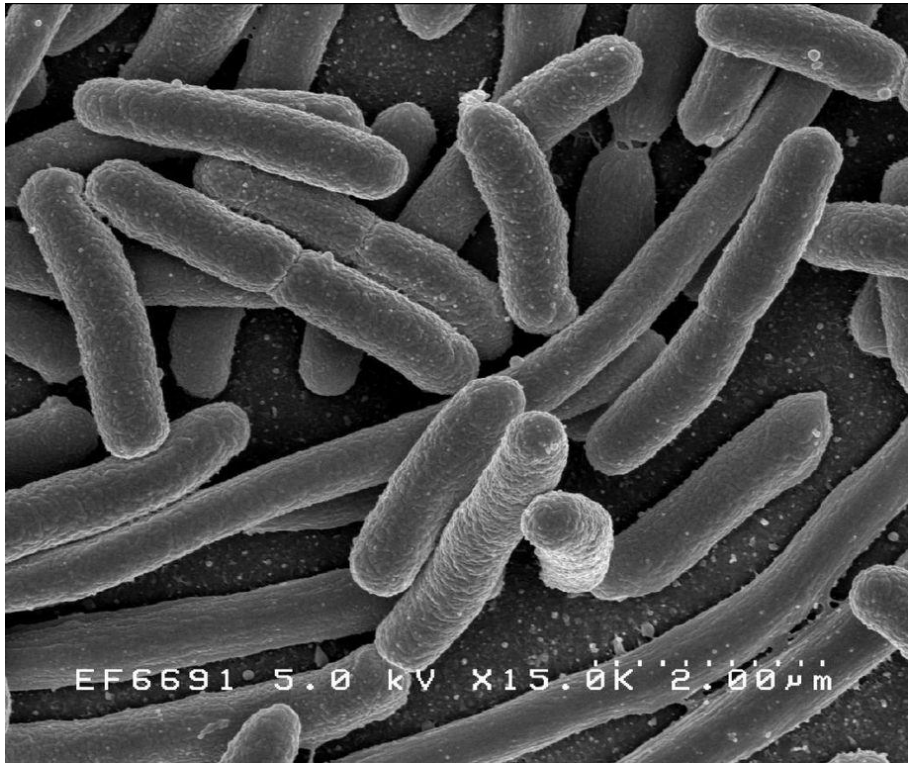
Professor, Seoul National University

CEO, ChunLab, Inc.

Version FEB-06-2019

www.chunlab.com

All lectures are available at
https://help.ezbiocloud.net/chun_lecture_kor/



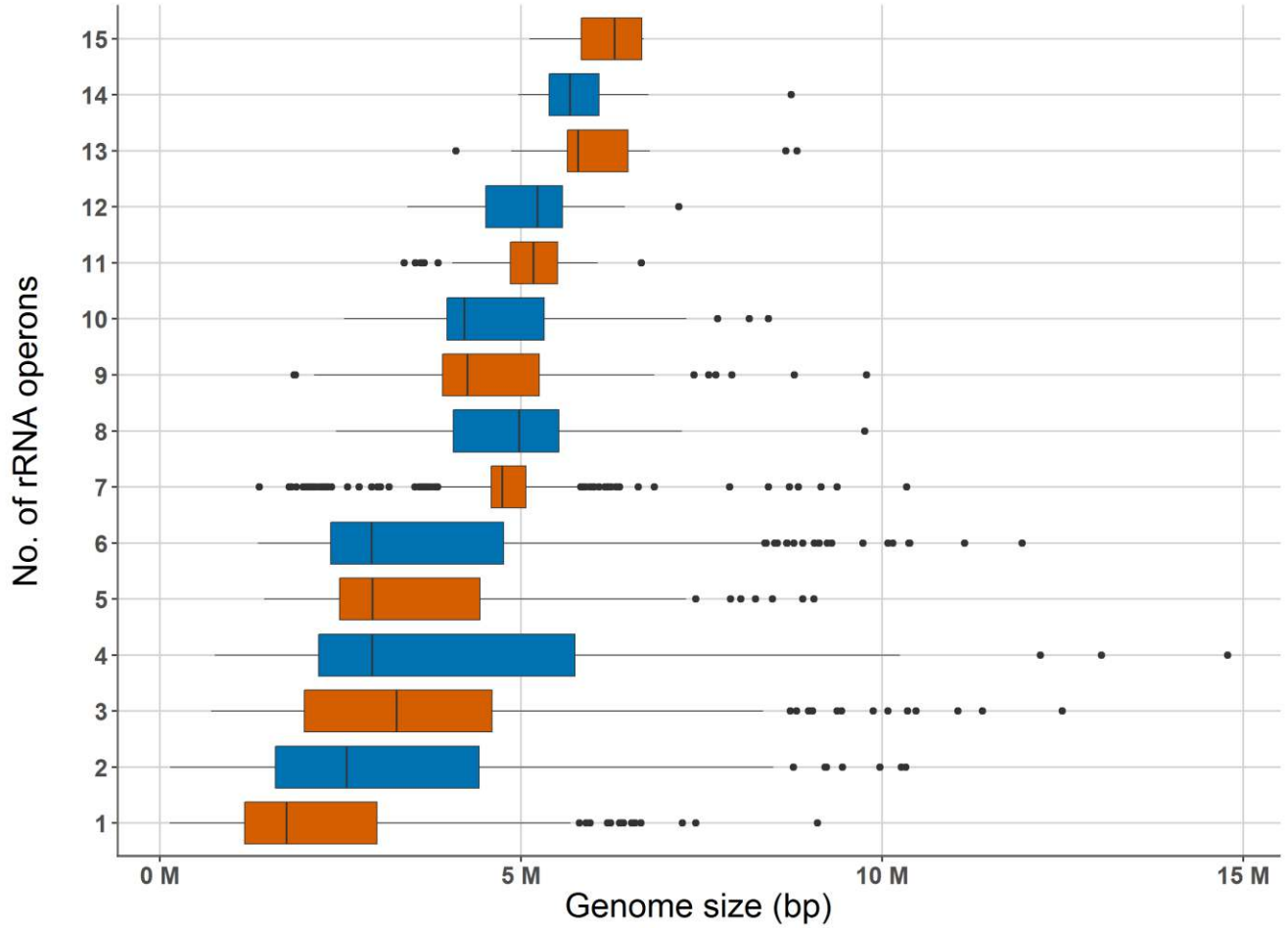
Wikipedia

Escherichia coli

https://www.ezbiocloud.net/genome/explore?puid=44

Based on the EzBioCloud database (4,131 complete genomes)

Number of rRNA operons vs. Genome size



Akkermansia muciniphila 1,000 reads

Bacteroides fragilis 1,000 reads

1:1 ratio ?

16S counts = cell counts ?

16S rRNA gene copy numbers of major gut microbiota

Akkermansia muciniphila	3 copies
Bacteroides fragilis	6 copies
Escherichia coli	7 copies
Faecalibacterium prausnitzii	6 copies
Prevotella copri	4 copies

All data available from https://www.ezbiocloud.net/

3 copies

Akkermansia muciniphila **1,000 reads**

6 copies

Bacteroides fragilis **1,000 reads**

1:1 ratio ?

3 copies

Akkermansia muciniphila **1,000 reads**

2:1 ratio

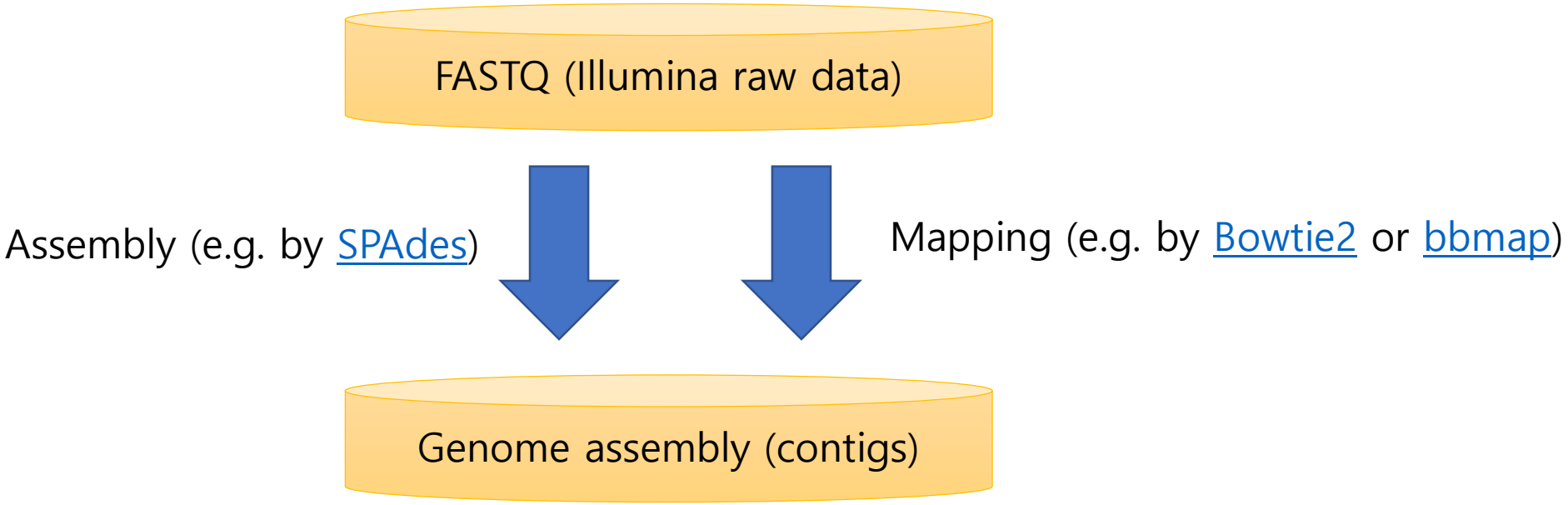
6 copies

Bacteroides fragilis **1,000 reads**

Method for 16S copy number correction

- **Genomes have the same 16S copy number if the strain belongs to the same species.**
- **How to obtain 16S copy number data from each species**
  1. From the complete genome assembly
     - » https://www.ezbiocloud.net/genome/explore?puid=648
  2. Using sequencing depth of 16S gene

Obtaining 16S copy number from Illumina raw data

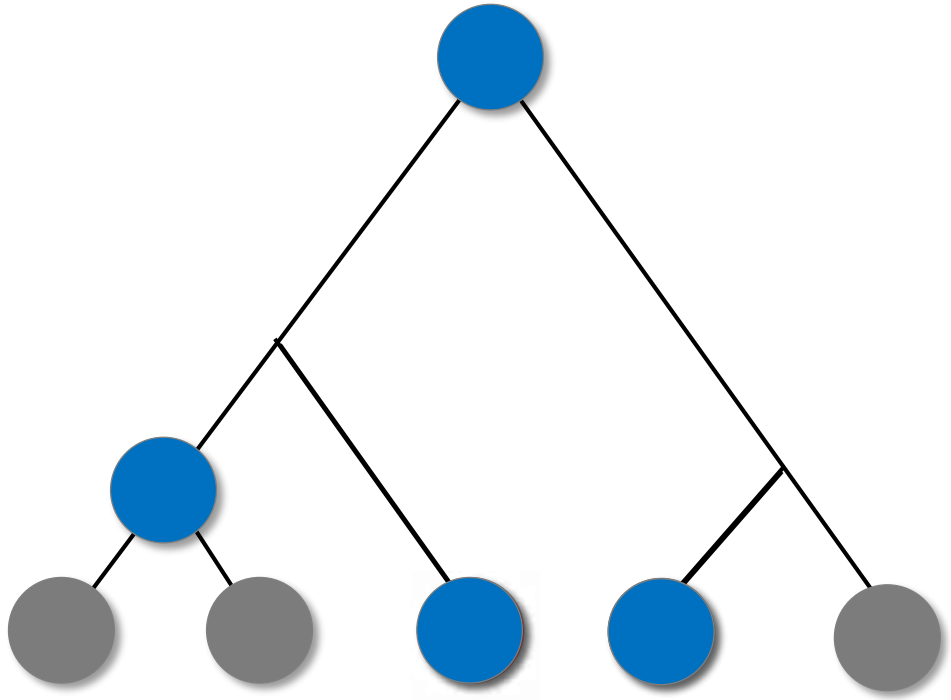


$$16S\ gene\ copy\ number = \frac{Sequencing\ depth\ of\ 16S\ rRNA\ gene}{Sequencing\ depth\ of\ whole\ genome}$$

Method for 16S copy number correction

- **Same copy number if the strain belongs to the same species.**
- **Getting copy number data from each species**
  1. From the complete genome assembly
     - » https://www.ezbiocloud.net/genome/explore?puid=648
  2. Using sequencing depth of 16S gene
  3. By predicting missing data

Predicting 16S copy numbers for missing data in the taxonomic tree



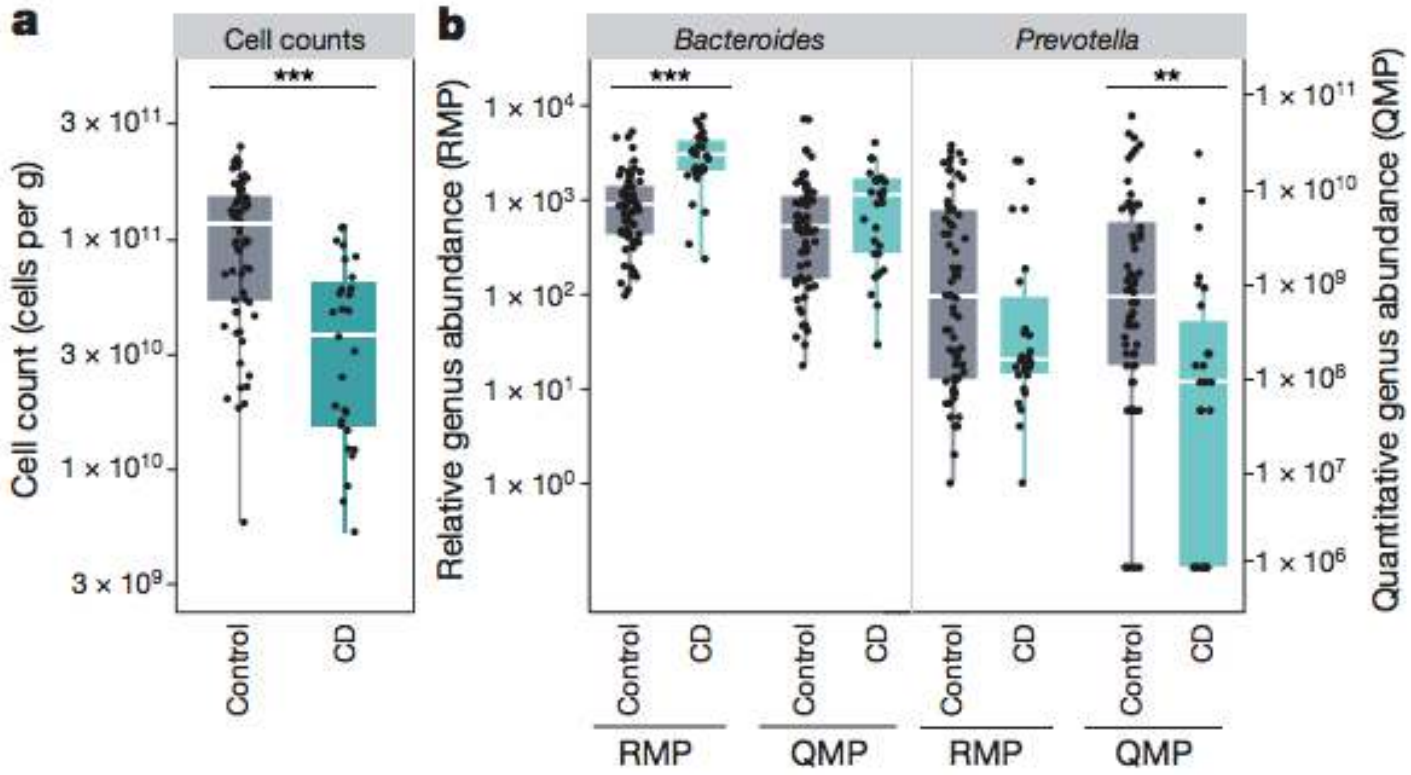
Node with data

Node without data

Node with inferred data

Calculated by **Ancestral State Reconstruction (ASR)** software in the PICRUSt package (Nat. Biotech. 31, 814–821; 2013)

Compositional vs. Corrected vs. Quantitative



Nature. 2017 Nov 23;551(7681):507-511.

LETTER

doi:10.1038/nature24460

Quantitative microbiome profiling links gut community variation to microbial load

Doris Vandeputte[1,2,3]*, Gunter Kathagen[1,2]*, Kevin D'hoe[1,2,3]*, Sara Vieira-Silva[1,2]*, Mireia Valles-Colomer[1,2], João Sabino[4], Jun Wang[1,2], Raul Y. Tito[1,2,3], Lindsey De Commer[1], Youssef Darzi[1,2], Séverine Vermeire[4], Gwen Falony[1,2]§ & Jeroen Raes[1,2]§

Figure 4 | Quantitative microbiome alterations in Crohn's disease. Microbiota alterations in a cohort of patients with Crohn's disease, using a relative or quantitative microbiome profiling approach. **a**, Differences in microbial load (cells per gram of faeces, log-transformed *y* axis) between healthy controls (Control, $n = 66$) and patients with Crohn's disease (CD, $n = 29$). Wilcoxon rank-sum test, $***P < 0.001$. **b**, Example of discordant genera when microbiota alterations in samples from patients with Crohn's disease are assessed using RMP or QMP (log-transformed *y* axis): *Bacteroides* is significantly increased in Crohn's disease patients using RMP, but the signal is lost using QMP. Conversely, *Prevotella* is detected as decreased in patients with Crohn's disease only when using QMP. The body of the box plot represents the first and third quartiles of the distribution and the median line. The whiskers extend from the quartiles to the last data point within 1.5× interquartile range, with outliers beyond. Kruskal–Wallis test, $***FDR < 0.001$, $**FDR < 0.01$.

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Incorporating 16S Gene Copy Number Information Improves Estimates of Microbial Diversity and Abundance

Steven W. Kembel[1¤]*, Martin Wu[2], Jonathan A. Eisen[3,4,5], Jessica L. Green[1,6]

1 Institute of Ecology & Evolution, University of Oregon, Eugene, Oregon, United States of America, **2** Department of Biology, University of Virginia, Charlottesville, Virginia, United States of America, **3** UC Davis Genome Center, University of California Davis, Davis, California, United States of America, **4** Department of Evolution and Ecology, College of Biological Sciences, University of California Davis, Davis, California, United States of America, **5** Department of Medical Microbiology and Immunology, School of Medicine, University of California Davis, Davis, California, United States of America, **6** Santa Fe Institute, Santa Fe, New Mexico, United States of America

Abstract

The abundance of different SSU rRNA ("16S") gene sequences in environmental samples is widely used in studies of microbial ecology as a measure of microbial community structure and diversity. However, the genomic copy number of the 16S gene varies greatly – from one in many species to up to 15 in some bacteria and to hundreds in some microbial eukaryotes. As a result of this variation the relative abundance of 16S genes in environmental samples can be attributed both to variation in the relative abundance of different organisms, and to variation in genomic 16S copy number among those organisms. Despite this fact, many studies assume that the abundance of 16S gene sequences is a surrogate measure of the relative abundance of the organisms containing those sequences. Here we present a method that uses data on sequences and genomic copy number of 16S genes along with phylogenetic placement and ancestral state estimation to estimate organismal abundances from environmental DNA sequence data. We use theory and simulations to demonstrate that 16S genomic copy number can be accurately estimated from the short reads typically obtained from high-throughput environmental sequencing of the 16S gene, and that organismal abundances in microbial communities are more strongly correlated with estimated abundances obtained from our method than with gene abundances. We re-analyze several published empirical data sets and demonstrate that the use of gene abundance versus estimated organismal abundance can lead to different inferences about community diversity and structure and the identity of the dominant taxa in microbial communities. Our approach will allow microbial ecologists to make more accurate inferences about microbial diversity and abundance based on 16S sequence data.

Citation: Kembel SW, Wu M, Eisen JA, Green JL (2012) Incorporating 16S Gene Copy Number Information Improves Estimates of Microbial Diversity and Abundance. PLoS Comput Biol 8(10): e1002743. doi:10.1371/journal.pcbi.1002743

Editor: Christian von Mering, University of Zurich and Swiss Institute of Bioinformatics, Switzerland

Received April 25, 2012; **Accepted** August 31, 2012; **Published** October 25, 2012

Funding: This research was supported by grant #1660 from the Gordon and Betty Moore Foundation. 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: kembel.steven_w@uqam.ca

¤ Current address: Département des sciences biologiques, Université du Québec à Montréal, Montreal, Quebec, Canada.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3486904/

Louca *et al. Microbiome* (2018) 6:41
https://doi.org/10.1186/s40168-018-0420-9

Microbiome

SHORT REPORT Open Access

Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem

Stilianos Louca[1,2*], Michael Doebeli[1,2,3] and Laura Wegener Parfrey[1,2,4]

"We recommend against correcting for 16S GCNs in microbiome surveys by default, unless OTUs are sufficiently closely related to sequenced genomes or unless a need for true OTU proportions warrants the additional noise introduced, so that community profiles remain interpretable and comparable between studies."

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5828423/

Genome database for predictions from 16S microbiome profiling

	EzBioCloud (As of 2018)	PICRUSt	tax4fun
No. of KEGG orthologs included	16,044	6,885	6,977
No. of genomes that were used for predicting KEGG orthologs	10,462	2,590	1,943
No. of genomes that were used for predicting 16S copy numbers	**3,188**	1,434	1,943

We will do both using the EzBioCloud!

Simulated dataset

Set 1

3 copies
Akkermansia muciniphila **200 reads**

6 copies
Bacteroides fragilis **200 reads**

Set 2

1 copies
Mycobacterium tuberculosis **200 reads**

7 copies
Escherichia coli **200 reads**

https://www.ezbiocloud.net/

Yi et al. BMC Infectious Diseases 2014, 14:583
http://www.biomedcentral.com/1471-2334/14/583

BMC Infectious Diseases

RESEARCH ARTICLE Open Access

Profiling bacterial community in upper respiratory tracts

Hana Yi[1,2,3], Dongeun Yong[4], Kyungwon Lee[4], Yong-Joon Cho[5] and Jongsik Chun[5,6*]

Abstract

Background: Infection by pathogenic viruses results in rapid epithelial damage and significantly impacts on the condition of the upper respiratory tract, thus the effects of viral infection may induce changes in microbiota. Thus, we aimed to define the healthy microbiota and the viral pathogen-affected microbiota in the upper respiratory tract. In addition, any association between the type of viral agent and the resultant microbiota profile was assessed.

Methods: We analyzed the upper respiratory tract bacterial content of 57 healthy asymptomatic people (17 health-care workers and 40 community people) and 59 patients acutely infected with influenza, parainfluenza, rhino, respiratory syncytial, corona, adeno, or metapneumo viruses using culture-independent pyrosequencing.

Results: The healthy subjects harbored primarily *Streptococcus*, whereas the patients showed an enrichment of *Haemophilus* or *Moraxella*. Quantifying the similarities between bacterial populations by using Fast UniFrac analysis indicated that bacterial profiles were apparently divisible into 6 oropharyngeal types in the tested subjects. The oropharyngeal types were not associated with the type of viruses, but were rather linked to the age of the subjects. *Moraxella nonliquefaciens* exhibited unprecedentedly high abundance in young subjects aged <6 years. The genome of *M. nonliquefaciens* was found to encode various proteins that may play roles in pathogenesis.

Conclusions: This study identified 6 oropharyngeal microbiome types. No virus-specific bacterial profile was discovered, but comparative analysis of healthy adults and patients identified a bacterium specific to young patients, *M. nonliquefaciens*.

Keywords: Microbiome, Respiratory tract, *Moraxella*, Influenza, Oropharynx, Healthcare staff

This dataset is provided as a tutorial set at
https://www.ezbiocloud.net/contents/16smtp
(Detailed document at
https://help.ezbiocloud.net/tutorial-mtp-respiratory-tract/)

Summary of the Chapter

- 16S rRNA gene copy numbers vary among the bacterial species.
- The compositions of cell counts can be predicted or inferred from the compositions of 16S counts using the reference database containing 16S copy numbers of each species.