genes

MDPI

Article

The Influences of Bioinformatics Tools and Reference Databases in Analyzing the Human Oral Microbial Community

Maria A. Sierra [1], Qianhao Li [1], Smruti Pushalkar [1], Bidisha Paul [1], Tito A. Sandoval [2], Angela R. Kamer [1], Patricia Corby [1], Yuqi Guo [1], Ryan Richard Ruff [3], Alexander V. Alekseyenko [4], Xin Li [1] and Deepak Saxena [1,5,*]

1 Department of Basic Science, New York University College of Dentistry, New York, NY 10010, USA; mas9996@nyu.edu (M.A.S.); kylin.qhl@outlook.com (Q.L.); sp117@nyu.edu (S.P.); bp1618@nyu.edu (B.P.); angela.kamer@nyu.edu (A.R.K.); patricia.corby@nyu.edu (P.C.); yg701@nyu.edu (Y.G.); xl15@nyu.edu (X.L.)
2 Department of Obstetrics and Gynecology, Weill Cornell Medicine, New York, NY 10065, USA; tas2037@med.cornell.edu
3 Department of Epidemiology & Health Promotion, New York University College of Dentistry, New York, NY 10010, USA; ryan.ruff@nyu.edu
4 The Biomedical Informatics Center, Program for Human Microbiome Research, Department of Public Health Sciences, Department of Oral Health Sciences, Department of Healthcare Leadership and Management, Medical University of South Carolina, Charleston, SC 29425, USA; alekseye@musc.edu
5 S. Arthur Localio Laboratory, Departments of Surgery New York University School of Medicine, New York, NY 10016, USA
* Correspondence: ds100@nyu.edu; Tel.: +1-212-9989256

Received: 6 June 2020; Accepted: 29 July 2020; Published: 3 August 2020

Abstract: There is currently no criterion to select appropriate bioinformatics tools and reference databases for analysis of 16S rRNA amplicon data in the human oral microbiome. Our study aims to determine the influence of multiple tools and reference databases on α-diversity measurements and β-diversity comparisons analyzing the human oral microbiome. We compared the results of taxonomical classification by Greengenes, the Human Oral Microbiome Database (HOMD), National Center for Biotechnology Information (NCBI) 16S, SILVA, and the Ribosomal Database Project (RDP) using Quantitative Insights Into Microbial Ecology (QIIME) and the Divisive Amplicon Denoising Algorithm (DADA2). There were 15 phyla present in all of the analyses, four phyla exclusive to certain databases, and different numbers of genera were identified in each database. Common genera found in the oral microbiome, such as *Veillonella*, *Rothia*, and *Prevotella*, are annotated by all databases; however, less common genera, such as *Bulleidia* and *Paludibacter*, are only annotated by large databases, such as Greengenes. Our results indicate that using different reference databases in 16S rRNA amplicon data analysis could lead to different taxonomic compositions, especially at genus level. There are a variety of databases available, but there are no defined criteria for data curation and validation of annotations, which can affect the accuracy and reproducibility of results, making it difficult to compare data across studies.

Keywords: 16S rRNA; databases; Greengenes; HOMD; NCBI; SILVA; RDP; QIIME; DADA2

1. Introduction

With decreasing costs, speed improvements, and throughput of DNA-sequencing techniques, analyses using marker genes (e.g., 16S rRNA or 18S rRNA) have become one of the most common methods for studying microbial communities [1]. The human body hosts an abundant and complex diversity of microbial communities, denominated the microbiome [2]. The human microbiome has

proven to be important in maintaining health, whereas dysbiosis is associated with various diseases and conditions [3]. Next-generation sequencing (NGS) technologies allow researchers to identify microbial taxa [4] and explore the possible role of the microbiomes in the human body [5].

Despite the wide use of 16S rRNA sequencing due to the latest advancements and benefits, errors and biases are introduced at different steps of the molecular experiment stage, from DNA extraction to sequencing, including amplification bias [6], chimeras [7], and biases introduced during computational analysis, such as Operational Taxonomic Unit (OTU) generation strategy, reference taxonomic sets, clustering algorithms, and specific software implementation [8,9]. Altogether, these methodologic differences could have dramatic effects on the accuracy of taxonomic classification, and α- and β-diversity estimation in 16S sequencing.

There are multiple bioinformatic tools available to analyze 16S rRNA gene amplicon sequencing data [10]. However, Quantitative Insights Into Microbial Ecology (QIIME) [11] and the Divisive Amplicon Denoising Algorithm (DADA2) [12] are among the most utilized [13]. Both pipelines are self-contained and can analyze 16S rRNA gene sequencing data from raw sequences (i.e., FASTQ), but they differ on how they cluster sequences: QIIME uses Operational Taxonomic Units (OTUs), sequences clustered with a fixed 3% dissimilarity threshold that might avoid fine-scale variation among sequences [14]. This method is used by most of the available pipelines [11,15–18]. Instead, DADA2 uses Amplicon Sequence Variants (ASV), an alternative error-modeling approach for denoising and clustering amplicons. Both pipelines enable the comparison of multiple and customized reference databases.

Taxonomic assignment is a crucial step in analyses. Reference databases are essential in the analysis of microbiomes because they are used to transform sequences into readable bacterial names. Reference databases for 16S taxonomy assignment include Greengenes [19], SILVA [20], the Ribosomal Database Project (RDP) [21], and the National Center for Biotechnology Information (NCBI) [22]). However, taxonomy assignment based on different reference databases might lead to different results [23].

The SILVA database contains information for all three domains of life (Bacteria, Archaea, and Eukarya). It is based on phylogenies for small subunit rRNAs (16S and 18S), and its taxonomic rank assignment is manually curated [24]. RDP database (Ribosomal Database Project) also contains rRNA sequences from the three domains, and most of the sequences are obtained from the International Nucleotide Sequence Database Collaboration (INSDC) [25]. Greengenes is a chimera-checked database that has Bacteria and Archaea sequences and most of the sequences are retrieved from the National Center for Biotechnology Information (NCBI) [26]. The NCBI taxonomy database contains the names of all organisms associated with submissions to the NCBI global database and is manually curated [22].

The human mouth harbors one of the most diverse and complex microbiomes in the human body, where up to 10,000 bacterial species have been identified [27]. Commensal taxa are associated with the development of dental caries and periodontal diseases [28], but it has also been associated with a higher risk of certain types of cancer (i.e., oral, pancreatic, gastrointestinal) [29–31]. Therefore, the standardization of bioinformatic tools and taxonomic reference databases for the analysis of the oral microbiome is key to correct taxa annotation to better understand the roles of oral microbiota in human health and disease. Here, we analyze the oral microbiome from 40 human saliva samples through two bioinformatic pipelines (i.e., QIIME and DADA2) and five reference databases (i.e., NCBI, Greengenes, SILVA, and RDP), including the Human Oral Microbiome Database (HOMD), which provides taxonomy for bacteria present in the human aerodigestive tract, including the oral cavity, pharynx, nasal passages, sinuses, and esophagus [32].

2. Materials and Methods

2.1. Collection of Samples

We collected saliva samples from 40 e-cigarette users. The Institutional Review Board of New York University Langone Medical Center approved the study, and all of the subjects provided informed consent and completed the questionnaires. Project identification code i16-00124: Impact of E-cigarette

on Oral Health, approved on 3/2/2016. All subjects were initially screened for their carbon monoxide (CO) levels by an exhaled CO breath test (Smokerlyzer, Covita, Santa Barbara, CA, USA) and salivary cotinine levels using test strips (NicAlert, Craig Medical Inc., Vista, CA, USA) [33].

Periodontal health status was determined by a comprehensive oral examination and only subjects with mild to severe periodontal disease were included in the study. The inclusion criteria were as follows: aged at least 21 years; systemically healthy, as evidenced by medical history; and currently using e-cigarettes (never smoked and using 0.5 to 1 e-cig/day for past 6 months). They were diagnosed with mild, moderate, or severe periodontal disease, according to the CDC in collaboration with the American Academy of Periodontology (CDC-AAP) [34].

The exclusion criteria were: having a medical condition (including uncontrolled diabetes and HIV); subjects who reported taking antibiotics or having a professional dental cleaning within 1 month of the enrollment day; a recent febrile illness that delayed or precluded participation; pregnancy/lactation; enrollment in other studies; a history of radiation therapy to the head and neck region; the presence of oral mucosal lesions suspected of candidiasis; herpes labialis; aphthous stomatitis; and premalignancy/malignancy, such as leukoplakia or erythroplakias. The participants were asked to chew paraffin wax pellets (Gleegum, Verve Inc., Providence, RI) to stimulate salivary flow rate, and saliva samples were collected for 5 min. The saliva was aliquoted to a desired volume and stored at −80 °C until further analysis.

2.2. DNA Extraction and Sequencing

Genomic DNA was extracted from saliva samples using the MoBio Power fecal kit, following the manufacturer's instructions (MoBio Laboratories Inc., Carlsbad, CA, USA). DNA was quantified for concentration and purity using a NanoDrop 2000 spectrophotometer (Thermo Scientific, Waltham, MA, USA) and stored at −20 °C until further analysis.

For high-throughput 16S library preparation and sequencing, the V3–V4 region of the 16S rRNA gene was amplified from the genomic DNA of saliva samples, according to the modified Illumina 16S metagenomics protocol (Part # 15,044,223 Rev. B). The purified DNA was quantified using the Quant-iT PicoGreen assay (Molecular Probes, Inc., Eugene, OR, USA) in a SpectraMax M5 microplate reader (Molecular Devices, Sunnyvale, CA, USA), and the concentrations were adjusted to 10 ng/µL for all sequencing assays. PCR was initially performed using the primer set, 341F (5′-CCTACGGGNGGCWGCAG-3′) and 805R (5′-GACTACHVGGGTATCTAATCC-3′), each with overhang adapter sequences (IDT, Coralville, IA, USA) using 2× Kapa HiFi Hotstart ReadyMix DNA polymerase (KapaBiosystems, Wilmington, MA, USA).

Samples were amplified in duplicates and purified using AMPure XP beads. Amplification was performed at 95 °C (3 min) with 25 cycles of 95 °C (30 s), 55 °C (30 s), 72 °C (30 s), and a final extension of 72 °C (5 min). Dual indices from Illumina Nextera XT index kits (Illumina, San Diego, CA) were added to target amplicons in a second PCR using 2× Kapa HiFi Hotstart ReadyMix DNA polymerase. PCR conditions were 95 °C (3 min), followed with 8 cycles of 95 °C (30 s), 55 °C (30 s), 72 °C (30 s), and a final extension of 72 °C (5 min). After each PCR cycle, AMPure XP bead-purified libraries were checked for purity using NanoDrop, quantified using PicoGreen assay, and size-confirmed on agarose gels. Negative controls were included in all sequencing runs. Equimolar amounts of the generated libraries were combined and quantified. The pooled amplicon library was denatured, diluted, and sequenced on an Illumina MiSeq platform using MiSeq Reagent Kit v3 (600 cycles), following the 2 × 300-bp paired-end sequencing protocol.

2.3. Bioinformatic and Statistical Analysis

QIIME scripts were run on the NYU High Performance Computing Cluster (HPC). Quality control of sequences was performed with FastQC [35].

QIIME v1.9.1 was used to process the data. Cutadapt (v1.12) [36] was used to remove the primers from both forward and reverse sequences. The cleaned sequences were merged using

join_paired_ends.py, and barcodes were extracted from the sequence headers. The sequences were then pooled, de-multiplexed, and filtered for quality control using multiple_split_libraries_fastq.py. The open-reference OTU-picking method, pick_open_reference_otus.py, was used to create an OTU table with default settings. Sequences were aligned with parallel_align_seqs_pynast.py using the PyNAST default method. The Usearch61 method was used to perform de novo and reference-based chimera detection using the parallel_identify_chimeric_seqs.py, and the chimeric sequences were filtered out using filter_fasta.py. The alignment built with PyNAST was filtered to remove highly variable regions with filter_alignment.py, and the OTU table was filtered for chimeric OTUs with filter_otus_from_otu_table.py.

DADA2 v1.14 was used with the pipeline tutorial (available at https://github.com/benjjneb/dada2). Sequence reads were first filtered using DADA2's default parameters (i.e., an expected error threshold of 2 with trimming of 250 and 200 bases for forward and reverse, respectively). Filtered reads were then de-replicated and de-noised using DADA2 default parameters. De-replication combines identical reads into unique sequences and constructs consensus quality profiles for each combined lot of sequences [37]. The consensus quality profiles then inform the de-noising algorithm, which infers error rates from samples and removes identified sequencing errors from the samples.

The phyloseq (v1.27.0) [38] was used to format OTUs/ASVs tables and calculate diversities of samples. α-diversity measures, such as Richness (measured by Observed OTUs, Abundance-based Coverage Estimator (ACE), and Chao1) and Species evenness (measured by Shannon index), were calculated using phyloseq (v1.27) [38] and metagMisc (0.0.4) available is R. For α-diversity measures of QIIME output, no normalization was made. However, a log10 normalization was made on the total abundance at phylum level presented in Figure 1. Normality (Shapiro–Wilk Test) and one-way ANOVA were used to evaluate the significant differences in α-diversity measures in Prism 8 (GraphPad), with p-values ≤ 0.05 considered significant. For β-diversity comparisons, a prevalence heatmap of presence/absence at phylum level was built. Additionally, Non-metric Multidimensional Scaling (NMDS) was calculated at phylum and genus level using vegan (v2.5.4) [39] package in R (v3.6).

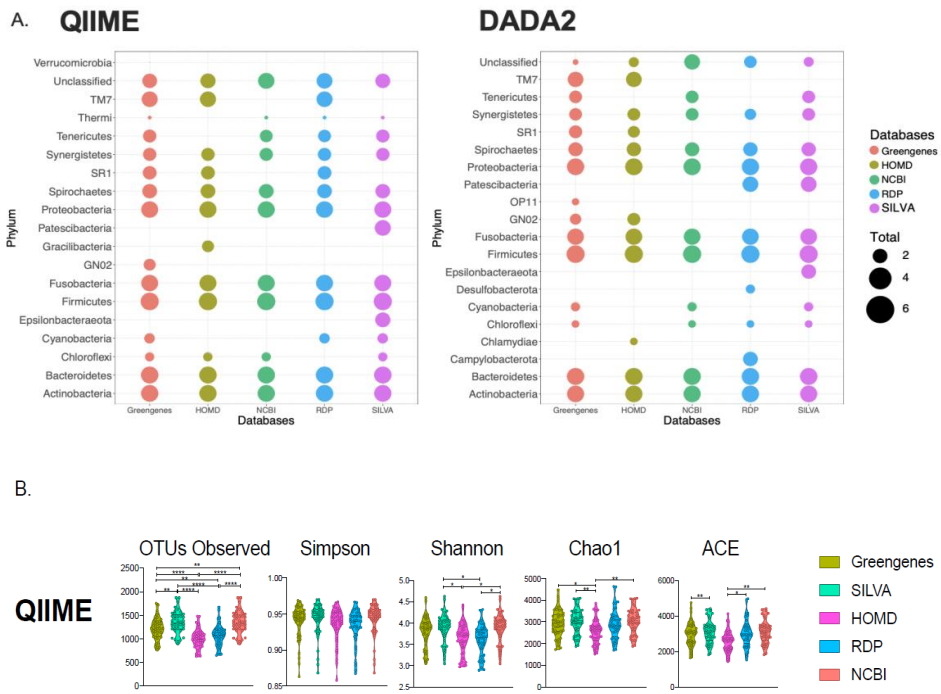

Figure 1. (**A**) Dotplot of phylum abundances from Quantitative Insights Into Microbial Ecology (QIIME) and the Divisive Amplicon Denoising Algorithm (DADA2) pipelines, comparing the five reference databases. Total abundances are log10 transformed. (**B**) α-diversity measurements for QIIME pipeline. p-values are assigned as ≤0.05 (*), <0.002 (**), <0.0002 (***), and <0.0001 (****).

2.4. Reference Databases

The FASTA files and taxonomy tables of the 16S rRNA gene for each database were downloaded from their respective websites.

Greengenes (v13_8) was downloaded from each pipeline website.

QIIME: http://qiime.org/home_static/dataFiles.html

DADA2: https://benjjneb.github.io/dada2/training.html

HOMD (v15.1, updated at 11/16/2017) was downloaded from the eHOMD website (http://www.homd.org), starting from position 9.

The NCBI-curated collection of 16S rRNA RefSeq sequences from the bacteria and archaea Targeted Loci Project was downloaded on January 28, 2019. "33175[BioProject] OR 33317[BioProject]" was used to search in the NCBI Nucleotide database. All of the items (21,075 in total) were downloaded as a FASTA file. The taxonomic information was extracted and downloaded using the tool, entrez_qiime, available on GitHub (https://github.com/bakerccm/entrez_qiime).

SILVA (v132, released at 04/10/2018) 16S-only reference sequences and taxonomy were downloaded from the SILVA website for QIIME and DADA2.

The RDP (v11) aligned Bacteria FASTA file and taxonomy file were downloaded from the webpage and used for both pipelines, QIIME and DADA2 (https://rdp.cme.msu.edu/misc/resources.jsp).

2.5. Data Availability

The sequences generated and analyzed during the current study are available in the NCBI repository, accession number PRJNA602902.

3. Results

A total of 5,453,541 sequences from 40 samples were processed using QIIME and DADA2. QIIME took approximately 4.5 h to produce a raw OTU table, while it took DADA2 1.5 h to produce a raw ASV table. By using QIIME, 3,125,624 sequences remained after merging and demultiplexing. These sequences were clustered into Greengenes: 16,018, HOMD: 14,291, NCBI: 16,028, SILVA: 16,078, and RDP: 16,426 OTUs. By using DADA2 and each of the databases, 2,750,305 sequences remained after quality control, denoising and merging, which were clustered into 9264 ASVs.

3.1. Comparisons of Taxonomic Composition and Diversity from Different Databases

Different phyla were detected when assigning taxonomic classification using different reference databases in QIIME (Table S1) and DADA2 (Table S2). Although there were 15 phyla present in all the analyses, four phyla were only present in certain databases, which was the case for the phylum candidatus Patescibacteria, detected in the SILVA database from the QIIME pipeline (Figure 1A), and the phylum Chlamydiae, which was only present in the HOMD database in the DADA2 pipeline. Phyla Actinobacteria, Fusobacteria, Firmicutes, Proteobacteria, and Synergistetes were among the most abundant phyla and annotated in all five databases. There were, however, multiple unclassified taxa, especially from the QIIME pipeline, where all the databases retrieved an average of ~38,000 unclassified OTUs, with a minimum of 11,892, from Greengenes, and a maximum of 98,058, from the NCBI database. α-diversity measures also displayed these differences in databases when analyzing with QIIME pipeline (Figure 1B). The number of observed OTUs in all databases showed statistically significant differences. Shannon index, Chao1, and ACE also were different. Since the DADA2 clustering method is based on the sequence variance, it does not produce singletons; thus, standard α-diversity measures could not be calculated.

Non-metric multidimensional scaling (NMDS) depicts the differences of databases at phylum level in both pipelines (Supplementary Figure S1A,B). In QIIME, most databases cluster together, except for the SILVA database, which clusters apart. While using DADA2, taxonomic assignment with the RDP database is more dissimilar than with other databases.

3.2. Comparison of Taxonomic Annotation at Genus Level

Using the DADA2 pipeline, a total of 128 different genera were identified in the samples by Greengenes, 119 by HOMD, 127 by NCBI, 158 by SILVA, and 146 by RDP databases. As for the QIIME pipeline, 217 genera were retrieved with SILVA, 186 with RDP, 207 with NCBI, 119 with HOMD, and 175 with Greengenes. A prevalence heatmap displays the presence of approximately 30 genera in all databases in both pipelines (Figure 2A,B). These common taxa include genera such as *Veillonella*, *Rothia*, and *Prevotella*, among others. Some genera were only present in certain databases, such as *Bulleidia* and *Paludibacter*, which were only found when using the Greengenes database, and the genera, *Ralstonia* and *Aerococcus*, which were only annotated when using the HOMD database and the DADA2 pipeline. Depending on the database, some genera are named with numbers, which corresponds to isolated strains. This is the case for *TG5*, present in Greengenes and assigned to phylum Synergistetes, and *F0058*, present in SILVA and assigned to Bacteroidetes. These taxa have been related to oral microbiome studies and associated with dental disease [40,41].

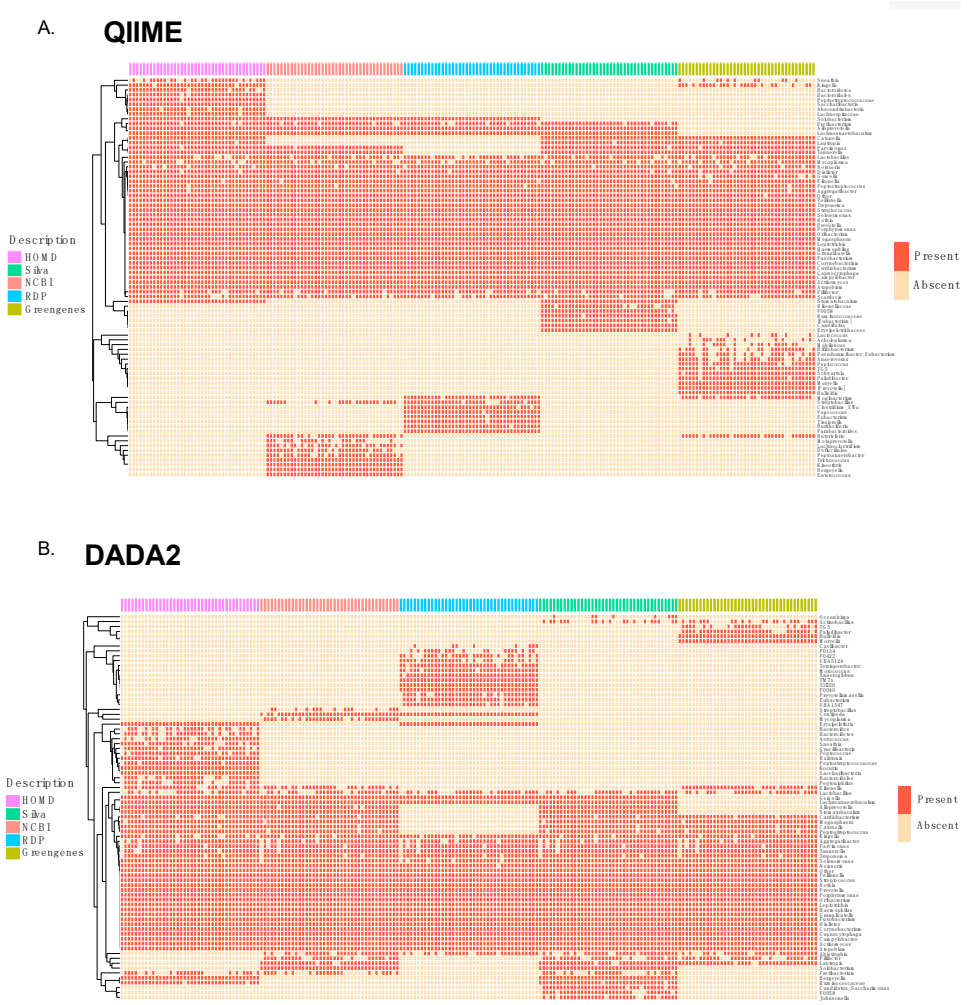

Figure 2. Prevalence heatmap of presence/absence of the 50 most abundant genera in (**A**) QIIME and (**B**) DADA2 pipelines.

For each of the 50 most abundant genera, we built a tree for each of the two pipelines and calculated the cophenetic distance matrix (Figure 3A), measuring the distances between leaves of the phylogenetic tree through branch lengths, as implemented in the R package stats. Values of correlation were $r = 0.889$ and $r = 0.898$ for QIIME and DADA2, respectively. These differences were also evident when performing non-metric multidimensional scaling (NMDS) (Figure 3B). Using the QIIME pipeline,

we found a greater dissimilarity of sample annotation, and even more when choosing the SILVA or Greengenes databases, while there was an evident clustering of samples annotated with NCBI and RDP. On the contrary, samples analyzed with DADA2 seemed not to be as affected by the choice of Greengenes, NCBI, or SILVA databases; however, RDP clustered apart, showing a greater dissimilarity.

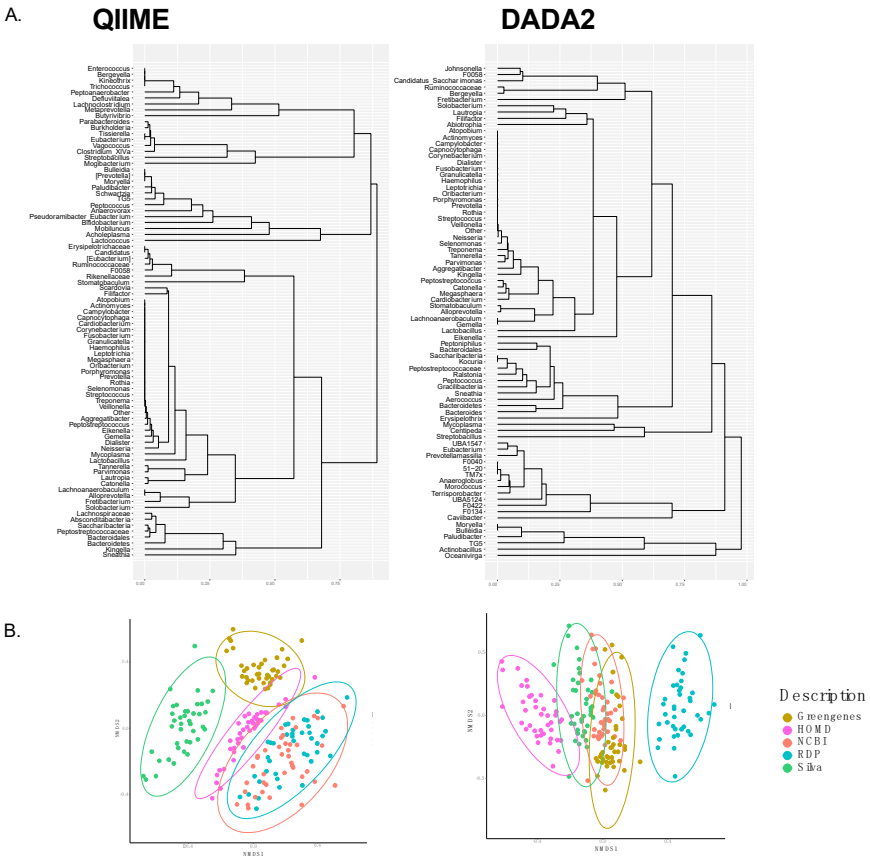

Figure 3. Comparisons of hierarchical clusters at genus level. (**A**) Cladogram of the 50 most abundant genera in each pipeline, with cophenetic correlation coefficients of $r = 0.889$ and $r = 0.898$ for QIIME and DADA2, respectively. (**B**) Non-metric multidimensional scaling (NMDS) at genus level, based on a Bray–Curtis dissimilarity matrix.

4. Discussion

Over the last few decades, next-generation sequencing (NGS) has greatly improved investigations into complex microbial communities. The development of computational methods has also played an essential role in this process by enabling researchers to analyze and transfer sequencing data into human-readable results. In this study, we analyzed 16S data from human saliva samples using five different reference databases and two different bioinformatics pipelines to compare their influence in exploring the composition of oral microbiomes. Fifteen phyla were commonly detected by the five databases, although Actinobacteria, Fusobacteria, Firmicutes, Proteobacteria, and Synergistetes were among the most abundant and prevalent in all databases. The dissimilarities shown in the NMDS in both pipelines using certain databases at phylum level might be due to, for QIIME, the SILVA database not including phyla TM7 and SR1, which seem highly abundant in these oral samples when using other reference databases. Additionally, SILVA includes the candidatus Patescibacteria and the phylum Epsilonbactereaota, which has been classified as an independent phylum in some studies, but as a Proteobacteria class in others [42]. For DADA, the dissimilarity in the RDP database might be because it classifies Firmicutes in three additional groups, Firmicutes_A, Firmicutes_B,

and Firmicutes_C, and does not include phyla TM7, GN02, or SR1, which are frequently detected in the human microbiome, including the human oral cavity [43].

Researchers should be cautious when choosing a particular database. Public databases might annotate multiple unverified phyla, as was the case with Campylobacterota and Desulfobacterota in RDP and Epsilonbactereota in SILVA, as well as create misannotations that could lead to false positives [44], which we suggest was the case with Verrucomicrobia, which was only annotated in two OTUs using Greengenes. Using custom databases and reducing the size of the reference database to the specific microbiome, one might avoid misannotations and improve taxonomic assignation at lower taxonomic levels [45]. However, a small reference database could skip certain taxa that have not yet been annotated, due to a lack of sufficient sequencing effort or a lack of previous isolation. As an example, the phylum Tenericutes was not annotated when using HOMD, but seemed highly present when using other databases.

When we compare the taxonomic annotations at genus level, half of the taxa were annotated independently of the database or the pipeline. These common taxa, which we denominated as "core-genera", seem to be commonly associated with human commensals, especially from the oral microbiome [46]. Nonetheless, some genera are excluded from HOMD and other customized databases, as is the case for *Stomatobaculum* [47], *Lactococcus* [48], *Bulleida* [49], and *Vagococcus* [50]. Other genera, such as *Acholeplasma*, *Mobiluncus*, *Anaerovorax*, and *Prevotellamassilia*, that have not been broadly associated with the oral microbiome, were also omitted in HOMD and only annotated in public reference databases, which could suggest that this customized database should be further nourished with new taxa annotations.

The differences found in the genera assignation were shown in the cladogram, cophenetic distance, and NMDS. This depicts the influence that a reference database might have on taxonomic assignment at genus level. These differences were not as notable when working with a higher phylogenetic level (i.e., phylum), although, when analyzing the role of the microbiome in human health and diseases, it is crucial to identify microbial taxa as precisely as possible (i.e., genus or species level) [51].

Extracting valuable information from enormous amounts of 16S sequencing data requires not only high-quality reference sequences, but also accurate and validated annotations. As researchers continue to extensively use these open-source reference databases, improved quality control will be necessary during their curation and validation. NCBI might arise as a good option as a scaffold, due to its large-scale sources and daily updates. However, custom and up-to-date databases are encouraged in order to validate the veracity of taxonomic annotations, avoid misclassifications, and improve taxonomic assignation at lower taxonomic levels. Additionally, databases with validated taxonomy and phenotypic and ecological characterization of species found in the microbiome are essential to understanding the role of the microbes in the microbiome [52,53].

In our study, we also compared QIIME and DADA2 pipelines to analyze our dataset. In accordance with other studies [54], DADA2 identified fewer ASVs than the number of OTUs identified by QIIME. QIIME is used in most microbiome studies [55,56], and provides a Biological Observation Matrix (BIOM file), which is useful for a wide range of downstream analyses. However, DADA2 might be faster and easier for less-experienced users, and arises as an alternative error-modeling approach for denoising and clustering amplicons. Altogether, these data show that the reference database used to align sequences and assign taxonomy information can have an effect on the final results.

5. Conclusions

Our study showed that using different reference databases in 16S sequencing data analyses could lead to different taxonomic compositions, especially at genus level. Currently, there are a variety of databases available, but there are no defined criteria for data curation and validation of annotations. This could affect the accuracy and reproducibility of results, and also makes it difficult to compare data from other studies. A well-curated and up-to-date microbiome-specific database is needed to improve

the reliability of 16S sequencing analyses and taxonomic annotations. The HOMD database might be a good start, but our results suggest that more taxa should be included.

We also compared QIIME and DADA2 in our study and found that at phylum level the choice of pipeline might not affect the results as much as at a lower level (i.e., genus), which could have a bigger impact on the results. Nevertheless, most of the common oral genera were present in both pipelines. Even though the number of OTUs (QIIME) was bigger than the number of ASVs (DADA2), our results show that the choice of certain databases might significantly affect the output, rather than the choice of pipeline. It is critical for researchers to consider these differences with regard to the questions they seek to answer and the type of microbiome under study.

Supplementary Materials: The following are available online at http://www.mdpi.com/2073-4425/11/8/878/s1. Supplemental Figure S1: NMDS at phylum level of both pipelines based on a Bray–Curtis dissimilarity matrix. Supplemental Table S1: Phylum total abundance for each of the five databases in the QIIME pipeline. Supplemental Table S2: Phylum total abundance for each of the five databases in the DADA2 pipeline.

Author Contributions: Conception and design, X.L., Q.L., and D.S.; development of methodology, M.A.S., Q.L., S.P., P.C., A.R.K., B.P., Y.G., and D.S.; acquisition of data, M.A.S., Q.L., S.P., P.C., A.R.K., B.P., Y.G., and D.S.; analysis and interpretation of data, M.A.S., Q.L., T.A.S., and D.S.; writing, review, and revision of the manuscript, M.A.S., Q.L., T.A.S., S.P., R.R.R., A.V.A., and D.S.; administrative, technical, or material support, Q.L., S.P., and D.S.; study supervision, M.A.S., X.L., and D.S. All authors have read and approved the manuscript.

Funding: This research project is supported by NIH grants CA206105 (D.S.), DE025992 (D.S. and X.L.), and DE027074 (D.S. and X.L.) and the NYU Mega grant initiative (D.S. and X.L.). Funders had no role in study design and data analysis.

Acknowledgments: The authors would like to thank the NIH and NYU Mega grant initiative for their funding support for laboratory supplies, sampling, and data analysis. Also, we thank Rebeca Vasconcelos for her help on samples acquisition.

Conflicts of Interest: The authors declare that they have no competing interests.

References

1. Pollock, J.; Glendinning, L.; Wisedchanwet, T.; Watson, M. The Madness of Microbiome: Attempting To Find Consensus "Best Practice" for 16S Microbiome Studies. *Appl. Environ. Microbiol.* **2018**, *84*, 7. [CrossRef]
2. Turnbaugh, P.J.; Ley, R.E.; Hamady, M.; Fraser-Liggett, C.M.; Knight, R.; Gordon, J.I. The Human Microbiome Project. *Nature* **2007**, *449*, 804–810. [CrossRef]
3. Ranjan, R.; Rani, A.; Metwally, A.; McGee, H.S.; Perkins, D.L. Analysis of the Microbiome: Advantages of Whole Genome Shotgun Versus 16s Amplicon Sequencing. *Biochem. Biophys. Res. Commun.* **2016**, *469*, 967–977. [CrossRef]
4. Woo, P.C.; Lau, S.K.; Teng, J.L.; Tse, H.; Yuen, K.Y. Then and Now: Use of 16s Rdna Gene Sequencing for Bacterial Identification and Discovery of Novel Bacteria in Clinical Microbiology Laboratories. *Clin. Microbiol. Infect.* **2008**, *14*, 908–934. [CrossRef]
5. Belizario, J.E.; Napolitano, M. Human Microbiomes and their Roles in Dysbiosis, Common Diseases, and Novel Therapeutic Approaches. *Front. Microbiol.* **2015**, *6*. [CrossRef]
6. Sze, M.A.; Schloss, P.D. The Impact of DNA Polymerase and Number of Rounds of Amplification in PCR on 16S rRNA Gene Sequence Data. *mSphere* **2019**, *4*. [CrossRef]
7. D'Amore, R.; Ijaz, U.Z.; Schirmer, M.; Kenny, J.G.; Gregory, R.; Darby, A.C.; Shakya, M.; Podar, M.; Quince, C.; Hall, N. A Comprehensive Benchmarking Study of Protocols and Sequencing Platforms for 16S rRNA Community Profiling. *BMC Genom.* **2016**, *17*, 55. [CrossRef]
8. Golob, J.L.; Margolis, E.; Hoffman, N.G.; Fredricks, D.N. Evaluating the Accuracy of Amplicon-Based Microbiome Computational Pipelines on Simulated Human Gut Microbial Communities. *BMC Bioinform.* **2017**, *18*, 283. [CrossRef]
9. Edgar, R.C. Accuracy of Taxonomy Prediction for 16S rRNA and Fungal ITS Sequences. *Peer J.* **2018**, *6*, e4652. [CrossRef]
10. Plummer, E.; Twin, J.; Bulach, D.M.; Garland, S.M.; Tabrizi, S.N. A Comparison of Three Bioinformatics Pipelines for the Analysis of Preterm Gut Microbiota Using 16s rRNA Gene Sequencing Data. *J. Proteomics Bioinform.* **2015**, *8*, 283. [CrossRef]

11. Caporaso, J.G.; Kuczynski, J.; Stombaugh, J.; Bittinger, K.; Bushman, F.D.; Costello, E.K.; Fierer, N.; Pena, A.G.; Goodrich, J.K.; I Gordon, J.; et al. QIIME Allows Analysis of High-Throughput Community Sequencing Data. *Nat. Methods* **2010**, *7*, 335–336. [CrossRef]
12. Callahan, B.J.; McMurdie, P.J.; Rosen, M.J.; Han, A.W.; Johnson, A.J.; Holmes, S.P. DADA2: High-Resolution Sample Inference from Illumina Amplicon Data. *Nat. Methods* **2016**, *13*, 581–583. [CrossRef]
13. Nilakanta, H.; Drews, K.L.; Firrell, S.; Foulkes, M.; Jablonski, K.A. A Review of Software for Analyzing Molecular Sequences. *BMC Res. Notes* **2014**, *7*, 830. [CrossRef]
14. Rosen, M.J.; Callahan, B.J.; Fisher, D.S.; Holmes, S.P. Denoising PCR-Amplified Metagenome Data. *BMC Bioinform.* **2012**, *13*, 283. [CrossRef]
15. Schloss, P.D. Reintroducing Mothur: 10 Years Later. *Appl. Environ. Microbiol.* **2019**, *86*. [CrossRef]
16. Mysara, M.; Njima, M.; Leys, N.; Raes, J.; Monsieurs, P. From Reads to Operational Taxonomic Units: An Ensemble Processing Pipeline for Miseq Amplicon Sequencing Data. *Gigascience* **2017**, *6*, 1–10. [CrossRef]
17. Kumar, S.; Carlsen, T.; Mevik, B.-H.; Enger, P.; Blaalid, R.; Shalchian-Tabrizi, K.; Kauserud, H. CLOTU: An Online Pipeline for Processing and Clustering of 454 Amplicon Reads into Otus Followed by Taxonomic Annotation. *BMC Bioinform.* **2011**, *12*, 182. [CrossRef]
18. Hildebrand, F.; Tadeo, R.; Voigt, A.Y.; Bork, P.; Raes, J. LotuS: An Efficient and User-Friendly OTU Processing Pipeline. *Microbiome* **2014**, *2*, 30. [CrossRef]
19. McDonald, D.; Price, M.N.; Goodrich, J.; Nawrocki, E.P.; DeSantis, T.Z.; Probst, A.J.; Andersen, G.L.; Knight, R.; Hugenholtz, P. An Improved Greengenes Taxonomy with Explicit Ranks for Ecological and Evolutionary Analyses of Bacteria and Archaea. *ISME J.* **2011**, *6*, 610–618. [CrossRef]
20. Quast, C.; Pruesse, E.; Yilmaz, P.; Gerken, J.; Schweer, T.; Yarza, P.; Peplies, J.; Glockner, F.O. The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools. *Nucleic Acids Res.* **2013**, *41*, D590–D596. [CrossRef]
21. Maidak, B.L.; Cole, J.R.; Lilburn, T.G.; Parker, C.T.J.; Saxman, P.R.; Stredwick, J.M.; Garrity, G.M.; Li, B.; Olsen, G.J.; Pramanik, S.; et al. The RDP (Ribosomal Database Project) Continues. *Nucleic Acids Res.* **2000**, *28*, 173–174. [CrossRef]
22. Federhen, S. The NCBI Taxonomy Database. *Nucleic Acids Res.* **2012**, *40*, D136–D143. [CrossRef]
23. Balvociute, M.; Huson, D.H. SILVA, RDP, Greengenes, NCBI and OTT—How Do These Taxonomies Compare? *BMC Genom.* **2017**, *18*, 114. [CrossRef]
24. Yilmaz, P.; Parfrey, L.W.; Yarza, P.; Gerken, J.; Pruesse, E.; Quast, C.; Schweer, T.; Peplies, J.; Ludwig, W.; Glockner, F.O. The SILVA and "All-species Living Tree Project (LTP)" Taxonomic Frameworks. *Nucleic Acids Res.* **2014**, *42*, D643–D648. [CrossRef]
25. Nakamura, Y.; Cochrane, G.; Karsch-Mizrachi, I. International Nucleotide Sequence Database C: The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.* **2013**, *41*, D21–D24. [CrossRef]
26. DeSantis, T.Z.; Hugenholtz, P.; Larsen, N.; Rojas, M.; Brodie, E.L.; Keller, K.; Huber, T.; Dalevi, D.; Hu, P.; Andersen, G.L. Greengenes, A Chimera-Checked 16s rRNA Gene Database and Workbench Compatible with ARB. *Appl. Environ. Microbiol.* **2006**, *72*, 5069–5072. [CrossRef]
27. Dewhirst, F.E.; Chen, T.; Izard, J.; Paster, B.J.; Tanner, A.C.R.; Yu, W.-H.; Lakshmanan, A.; Wade, W.G. The Human Oral Microbiome. *J. Bacteriol.* **2010**, *192*, 5002–5017. [CrossRef]
28. Wade, W.G. The Oral Microbiome in Health and Disease. *Pharmacol. Res.* **2013**, *69*, 137–143. [CrossRef]
29. Ahn, J.; Chen, C.Y.; Hayes, R.B. Oral Microbiome and Oral and Gastrointestinal Cancer Risk. *Cancer Causes Controle* **2012**, *23*, 399–404. [CrossRef]
30. Gholizadeh, P.; Eslami, H.; Yousefi, M.; Asgharzadeh, M.; Aghazadeh, M.; Kafil, H.S. Role of Oral Microbiome on Oral Cancers, A Review. *Biomed. Pharmacother.* **2016**, *84*, 552–558. [CrossRef]
31. Fan, X.; Alekseyenko, A.V.; Wu, J.; Peters, B.A.; Jacobs, E.J.; Gapstur, S.M.; Purdue, M.P.; Abnet, C.C.; Stolzenberg-Solomon, R.; Miller, G.; et al. Human Oral Microbiome and Prospective Risk for Pancreatic Cancer: A Population-Based Nested Case-Control Study. *Gut* **2018**, *67*, 120–127. [CrossRef]
32. Escapa, I.F.; Chen, T.; Huang, Y.; Gajare, P.; Dewhirst, F.E.; Lemon, K.P. New Insights into Human Nostril Microbiome from the Expanded Human Oral Microbiome Database (eHOMD): A Resource for the Microbiome of the Human Aerodigestive Tract. *mSystems* **2018**, *3*. [CrossRef]

33. Marrone, G.F.; Paulpillai, M.; Evans, R.J.; Singleton, E.G.; Heishman, S.J. Breath Carbon Monoxide and Semiquantitative Saliva Cotinine as Biomarkers for Smoking. *Hum. Psychopharmacol. Clin. Exp.* **2010**, *25*, 80–83. [CrossRef]
34. Eke, P.I.; Dye, B.A.; Wei, L.; Slade, G.D.; Thornton-Evans, G.O.; Beck, J.D.; Taylor, G.W.; Borgnakke, W.S.; Page, R.C.; Genco, R.J. Self-Reported Measures for Surveillance of Periodontitis. *J. Dent. Res.* **2013**, *92*, 1041–1047. [CrossRef]
35. Brown, J.; Pirrung, M.; McCue, L.A. FQC Dashboard: Integrates FastQC Results into a Web-Based, Interactive, and Extensible Fastq Quality Control Tool. *Bioinformatics* **2017**, *33*, 3137–3139. [CrossRef]
36. Didion, J.P.; Martin, M.; Collins, F.S. Atropos: Specific, Sensitive, and Speedy Trimming of Sequencing Reads. *Peer J.* **2017**, *5*, e3720. [CrossRef]
37. Dahan, D.; Jude, B.A.; Lamendella, R.; Keesing, F.; Perron, G.G. Exposure to Arsenic Alters the Microbiome of Larval Zebrafish. *Front. Microbiol.* **2018**, *9*, 1323. [CrossRef]
38. McMurdie, P.J.; Holmes, S. Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE* **2013**, *8*, e61217. [CrossRef]
39. Dixon, P. VEGAN, A Package of R Functions for Community Ecology. *J. Vegetation Sci.* **2003**, *14*, 927–930. [CrossRef]
40. Dingsdag, S.; Nelson, S.; Coleman, N.V. Bacterial Communities Associated with Apical Periodontitis and Dental Implant Failure. *Microb. Ecol. Health Dis.* **2016**, *27*, 31307. [CrossRef]
41. SILVA rRNA Database Project. Available online: https://www.arb-silva.de/browser/lsu/CP002345 (accessed on 11 July 2020).
42. Trembath-Reichert, E.; Butterfield, D.A.; Huber, J.A. Active Subseafloor Microbial Communities from Mariana Back-Arc Venting Fluids Share Metabolic Strategies Across Different Thermal Niches and Taxa. *ISME J.* **2019**, *13*, 2264–2279. [CrossRef]
43. Baker, J.L.; Bor, B.; Agnello, M.; Shi, W.; He, X. Ecology of the Oral Microbiome: Beyond Bacteria. *Trends Microbiol.* **2017**, *25*, 362–374. [CrossRef]
44. Nobre, T.; Campos, M.D.; Lucic-Mercy, E.; Arnholdt-Schmitt, B. Misannotation Awareness: A Tale of Two Gene-Groups. *Front. Plant Sci.* **2016**, *7*, 868. [CrossRef]
45. Ritari, J.; Salojärvi, J.; Lahti, L.; de Vos, W.M. Improved Taxonomic Assignment of Human Intestinal 16S rRNA Sequences by a Dedicated Reference Database. *BMC Genomics* **2015**, *16*, 1056. [CrossRef]
46. Seshadri, R.; Myers, G.S.A.; Tettelin, H.; Eisen, J.A.; Heidelberg, J.F.; Dodson, R.J.; Davidsen, T.M.; DeBoy, R.T.; Fouts, D.E.; Haft, D.H.; et al. Comparison of the Genome of the Oral Pathogen *Treponema denticola* with Other Spirochete Genomes. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 5646–5651. [CrossRef]
47. Sizova, M.V.; Muller, P.; Panikov, N.; Mandalakis, M.; Hohmann, T.; Hazen, A.; Fowle, W.; Prozorov, T.; Bazylinski, D.A.; Epstein, S. Stomatobaculum Longum Gen. Nov., sp. Nov., an Obligately Anaerobic Bacterium from the Human Oral Cavity. *Int. J. Syst. Evol. Microbiol.* **2013**, *63*, 1450–1456. [CrossRef]
48. Robinson, K.; Chamberlain, L.M.; Schofield, K.M.; Wells, J.M.; Le Page, R.W. Oral Vaccination of Mice Against Tetanus with Recombinant Lactococcus Lactis. *Nat. Biotechnol.* **1997**, *15*, 653–657. [CrossRef] [PubMed]
49. Downes, J.; Olsvik, B.; Hiom, S.J.; Spratt, D.A.; Cheeseman, S.L.; Olsen, I.; Weightman, A.J.; Wade, W. Bulleidia extructa gen. nov., sp. nov., Isolated from the Oral Cavity. *Int. J. Syst. Evol. Microbiol.* **2000**, *50*, 979–983. [CrossRef]
50. Al-Ahmad, A.; Pelz, K.; Schirrmeister, J.F.; Hellwig, E.; Pukall, R. Characterization of the First Oral Vagococcus Isolate from a Root-Filled Tooth with Periradicular Lesions. *Curr. Microbiol.* **2008**, *57*, 235–238. [CrossRef]
51. Mason, M.R.; Nagaraja, H.N.; Camerlengo, T.; Joshi, V.M.; Kumar, P.S. Deep Sequencing Identifies Ethnicity-Specific Bacterial Signatures in the Oral Microbiome. *PLoS ONE* **2013**, *8*, e77287. [CrossRef] [PubMed]
52. Sierra, M.A.; Bhattacharya, C.; Ryon, K.; Meierovich, S.; Shaaban, H.; Westfall, D.; Mohammad, R.; Kuchin, K.; Afshinnekoo, E.; Danko, D.C.; et al. The Microbe Directory v2.0: An Expanded Database of Ecological and Phenotypical Features of Microbes. *BioRxiv* **2019**. [CrossRef]
53. Parks, D.H.; Chuvochina, M.; Waite, D.W.; Rinke, C.; Skarshewski, A.; Chaumeil, P.A.; Hugenholtz, P. A Standardized Bacterial Taxonomy Based on Genome Phylogeny Substantially Revises the Tree of Life. *Nat. Biotechnol.* **2018**, *36*, 996–1004. [CrossRef]

54. Allali, I.; Arnold, J.W.; Roach, J.; Cadenas, M.B.; Butz, N.; Hassan, H.M.; Koci, M.D.; Ballou, A.; Mendoza, M.; Ali, R.; et al. A Comparison of Sequencing Platforms and Bioinformatics Pipelines for Compositional Analysis of the Gut Microbiome. *BMC Microbiol.* **2017**, *17*, 194. [CrossRef]
55. Navas-Molina, J.A.; Peralta-Sánchez, J.M.; González, A.; McMurdie, P.J.; Vazquez-Baeza, Y.; Xu, Z.; Ursell, L.K.; Lauber, C.; Zhou, H.; Song, S.J.; et al. Chapter Nineteen—Advancing Our Understanding of the Human Microbiome Using QIIME. In *Methods in Enzymology*; DeLong, E.F., Ed.; Academic Press: Cambridge, MA, USA, 2013; Volume 531, pp. 371–444.
56. Yatsunenko, T.; Rey, F.E.; Manary, M.J.; Trehan, I.; Dominguez-Bello, M.G.; Contreras, M.; Magris, M.; Hidalgo, G.; Baldassano, R.N.; Anokhin, A.P.; et al. Human Gut Microbiome Viewed Across Age and Geography. *Nature* **2012**, *486*, 222–227. [CrossRef]