Journal of
Proteomics & Bioinformatics

Plummer et al., J Proteomics Bioinform 2015, 8:12
DOI: 10.4172/0974-276X.1000381

Research Article Open Access

A Comparison of Three Bioinformatics Pipelines for the Analysis of Preterm Gut Microbiota using 16S rRNA Gene Sequencing Data

Erica Plummer[1]*, Jimmy Twin[1], Dieter M. Bulach[2,3,4], Suzanne M. Garland[1,3,5,6] and Sepehr N Tabrizi[1,3,5,6]

1Murdoch Childrens Research Institute, The Royal Children's Hospital, Flemington Rd, Parkville, Victoria 3052 Australia
[2] Monash University, Victoria 3800, Australia
[3]The University of Melbourne, Victoria 3050, Australia
[4]Victorian Life Sciences Computation Initiative, The University of Melbourne, Parkville Campus, LAB-14, 700 Swanston St, Carlton, Victoria 3053, Australia
[5]The Royal Women's Hospital, 20 Flemington Rd, Parkville, Victoria 3052, Australia
[6]The Royal Children's Hospital, 50 Flemington Rd, Parkville, Victoria 3052, Australia

Abstract

Objective and Methods: Analysis of massive parallel sequencing 16S rRNA data requires the use of sophisticated bioinformatics pipelines. Several pipelines are available, however there is limited literature available comparing the features, advantages and disadvantages of each pipeline. This makes the choice of which method to use often unclear. Using gut microbial read data collected from a cohort of very preterm babies, we compared three pipelines commonly used for 16S rRNA gene analysis: MetaGenome Rapid Annotation using Subsystem Technology (MG-RAST), Quantitative Insights into Microbial Ecology (QIIME) and mothur. Using primarily default parameters, the three pipelines were compared in terms of taxonomic classification, diversity analysis and usability.

Results: Overall, the three pipelines detected the same phylum in similar abundances ($P>0.05$). A difference was observed between the pipelines in terms of taxonomic classification of genera from the *Enterobacteriaceae* family, specifically *Enterobacter* and *Klebsiella* ($P<0.0001$ and $P=0.0026$ respectively). We found the analysis time to be quickest with QIIME compared to mothur and MG-RAST (approximately 1 hour as compared to 10 hours and 2 days respectively).

Conclusion: This study showed that QIIME, mothur and MG-RAST produce comparable results and that regardless of which pipeline or algorithm is selected for the analysis of 16S rRNA gene sequencing data you are likely to generate a reliable high-level overview of sample composition when analysing faecal samples. The differences we observed at the genus level highlight that a key limitation of using 16S rRNA gene analysis for genus and species level classification is that related bacterial species may be indistinguishable due to near identical 16S rRNA gene sequences. This is important to keep in mind when analysing 16S rRNA gene sequencing data.

Keywords: Comparison; MG-RAST; QIIME; Mother; 16S rRNA gene analysis; Microbiome

Abbreviations: MG-RAST: MetaGenome Rapid Annotation using Subsystem Technology; QIIME: Quantitative Insights Into Microbial Ecology, W.A.T.E.R.s: A workflow for the alignment, taxonomy, and ecology of ribosomal sequences; RDPipeline: Ribosomal Database Project Pipeline; VAMPS; SnoWMan; OTU: operational taxonomic unit; RDP: Ribosomal Database Project; NCBI: National Center for Biotechnology Information; BLAT: BLAST like alignment tool; NMDS: non-metric multidimensional scaling; GUI: Graphical User Interface; M5RNA: Non-redundant multisource ribosomal RNA annotation; PyNAST: PythonNAST; MUSCLE: MUltiple Sequence Comparison by Log-Expectation; INFERNAL: INFERence of RNA Alignment; BLAST: Basic Local Alignment Search Tool; CD-HIT: Cluster Database at High Identity with Tolerance; PCA: Principal Coordinate Analysis

Introduction

Targeted amplicon based analysis using 16S rRNA gene sequences is commonly used to investigate complex bacterial communities such as the human gut [1]. Analysis of such communities requires the use of bioinformatics tools to efficiently and reproducibly process the large amount of data generated from amplicon sequencing to derive a taxonomic overview. There are various tools available to analyse 16S rRNA gene sequencing data including QIIME (Quantitative Insights Into Microbial Ecology) [2], mothur [3], MG-RAST (Metagenomics - Rapid Annotation using Subsystems Technology) [4], Genboree [5], EzTaxon [6], Pheonix2 [7], METAGENassist [8], MEGAN [9], VAMPS [10], SnoWMan [11], CloVR-16S [12], the RDPipeline (Ribosomal Database Project Pipeline) [13], Vegan [14], ade4 [15], and ape [16]. These tools can be categorised into those that are self-contained analysis pipelines i.e. those that incorporate various algorithms for quality control, clustering of similar sequences, assigning taxonomy, calculating diversity measures and visualising results and those that are not self-contained and can be used only for a specific step/s in the analysis of 16S rRNA gene sequencing data. There is limited literature comparing the functions and usability of these tools making the choice of which method to use often unclear.

In November 2014 Nilakanta et al. [17] published a review on the installation, documentation, features, and functions of seven tools: mothur, QIIME, W.A.T.E.R.S, RDPipeline, VAMPS, Genboree and SnoWMan. Nilakanta et al. concluded that mothur and QIIME were

***Corresponding authors:** Erica Plummer, Murdoch Childrens Research Institute, The Royal Children's Hospital, Flemington Rd, Parkville, Victoria 3052 Australia, Tel: +61 1300 766 439; E-mail: Erica.Plummer@mcri.edu.au

Received November 17, 2015; **Accepted** December 22, 2015; **Published** December 28, 2015

Citation: Plummer E, Twin J, Bulach DM, Garland SM, Tabrizi SN (2015) A Comparison of Three Bioinformatics Pipelines for the Analysis of Preterm Gut Microbiota using 16S rRNA Gene Sequencing Data. J Proteomics Bioinform 8: 283-291. doi:10.4172/0974-276X.1000381

the two outstanding pipelines due to their comprehensive features and support documentation. This review was limited in that it did not use a unified dataset to compare the performance of the seven pipelines in terms of taxonomic assignments generated. In 2014, D'Argenio et al. [18] compared taxonomic compositions and diversity measures generated by QIIME and MG-RAST using four gastrointestinal samples. The study found no statistically significant differences in the assignments or alpha diversity measures; however the study concluded that QIIME produced the more accurate assignments, primarily due to the high number of reads unable to be classified by MG-RAST. A statistically significant difference was observed between the two pipelines with regards to beta diversity measures and the authors hypothesised that this was likely due to the reads left unclassified by MG-RAST. This review was limited by the small number of samples included in the analysis.

In the current study, we use a single dataset (N=35) of human gut microbial read data collected from preterm infants to compare three commonly used pipelines, QIIME, mothur and MG-RAST, in terms of taxonomic assignments generated at both the phylum and genus level. We also compare and review the functionality and usability of QIIME, mothur and MG-RAST. We selected these three pipelines for comparison because despite having different workflows, they are all self-contained pipelines that can be used to analyse 16S rRNA gene sequencing data from start (i.e. raw sequence reads – SFF/FASTA/QUAL) to finish (generate OTU/abundance table), enable comparison of multiple samples and employ the use of the SILVA 16S rRNA gene reference database. Additionally despite being three of the most commonly used and cited pipelines for the analysis of 16S rRNA gene data, there is limited information available comparing these pipelines. The information obtained from this study will be of particular use and interest to researchers who are new to the field or who have limited bioinformatics experience as it provides an unbiased comparison and overview of three of the most commonly used bioinformatics pipelines for characterising bacterial communities using the 16S rRNA gene.

Methods

Samples

We used a subset of faecal samples collected from preterm infants who participated in The ProPrems Trial [19,20]. Participant infants were born at less than 32 weeks' gestation and weighed less than 1500 g; the infants were randomised to receive either a probiotic preparation (*Bifidobacterium longum* subsp. *Infantis* BB-02, *Streptococcus thermophilus* TH-4 and *B. animalis* subsp. *Lactis* BB-12) or a placebo. Up to seven longitudinal faecal swabs were collected from each infant recruited in Victoria, Australia, over the first 12 months of life (12 months corrected age). A total of 35 swabs from 15 infants were included in this study. Sample characteristics are shown in Table 1.

Ethics statements

The study has approval from The Royal Women's Hospital (Melbourne) Human Research Ethics Committee, and infant samples were collected after obtaining written informed consent from parents or guardians.

DNA extraction and sequencing

DNA was extracted from the samples using the MagNA Pure 96 System (Roche Diagnostics, Branchburg, NJ). The extracted DNA was used to generate an amplicon based library using *Bifidobacterium* optimised PCR primers 357F/926Rb (357F - CCTACGGGAGGCAGCAG, 926Rb – CCGTCAATT**Y**MTTTRAGT,

Infants, n	15
Specimen, n	35
Infants with 1 specimen, n	4
Infants with 2 specimen, n	4
Infants with 3 specimen, n	5
Infants with 4 specimen, n	2
Age at each specimen collection, d, mean (range)	102 (4-336)

Table 1: Relevant sample characteristics

the base that differs from the standard 926R primer is bolded) that target the V3-V5 hypervariable regions of the 16S rRNA gene as described by Sim et al. [21]. Each sample was barcoded with a Multiplex Identification (MID) tag (10 bp in length). The presences of tagged amplicons were confirmed visually on a 2% agarose gel. Amplicons were purified and pyrosequencing was performed on a Roche 454 Genome Sequencer (GS FLX Titanium Chemistry) at Macrogen Inc. (Seoul, South Korea).

Bioinformatics analysis

The resulting sequence read files were analysed using the three pipelines: QIIME (Version 1.8.0), mothur (Version 1.31.2) and MG-RAST (Version 3.3.7.3). The SILVA reference database [22] and a 97% similarity cut-off was used to cluster reads for taxonomic classification in each pipeline.

Chimera filtering was performed using UCHIME [23]. The rRNA 16S gold database [24] was used as the chimera checked reference database. In QIIME and MG-RAST, chimera filtering was done using UCHIME directly; we used the *chimera.uchime* command in mothur. A summary of the work flows used in each pipeline is shown in Figure 1. QIIME and mothur analysis was run on a Linux cluster (iDataplex × 86 system, Merri cluster at the Victorian Life Sciences Computational Initiative [25]).

QIIME

Reads were de-multiplexed and underwent quality control using the *split_libraries.py* command. The following quality control parameters were used: reads were removed if they were less than 250 bp in length, contained greater than eight ambiguous bases, contained homopolymers greater than eight base pairs in length, or had an average quality score of less than 25. No sliding window was used. OTU picking was performed using the *pick_otus.py* command with the default UCLUST algorithm.

The UCLUST algorithm uses the USEARCH algorithm to assign sequences to a cluster [26]. The USEARCH algorithm works by searching a query sequence against target sequences and recording the k-mers in common between the two sequences. Rather than inferring sequence similarity as the number of matching k-mers between a query and target sequence, USEARCH arranges the target sequences in decreasing order of the number of unique k-mers shared between the two sequences. The query sequences are arranged into clusters. Each cluster centroid shares a level of similarity below a set identity threshold level with each other centroid. The remaining query sequences are then assigned to a centroid (target sequence) based on identity threshold using the USEARCH algorithm described above. If the query sequence does not share similarity with a centroid above the threshold a new cluster is created.

The most abundant read in each OTU was selected as the representative sequence, this step was performed using *pick_rep_set.*

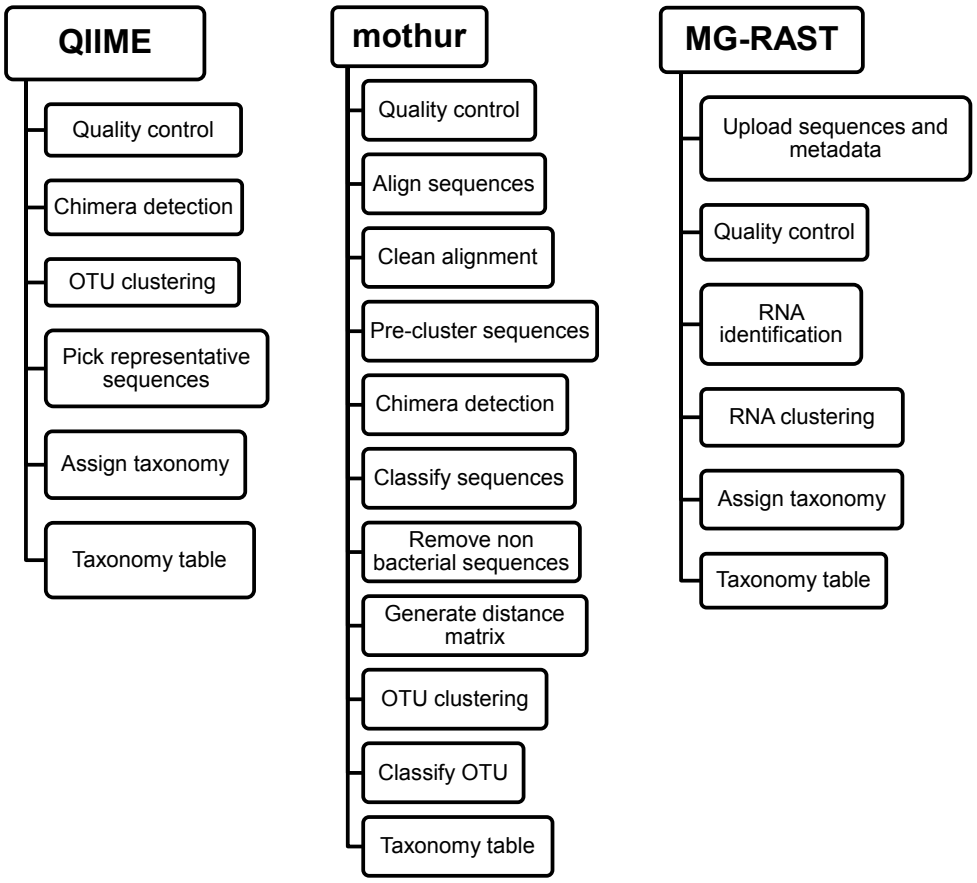

Figure 1: Overview of the workflows used by QIIME, mothur and MG-RAST. Several steps are shared across the three pipelines (e.g. quality control, clustering, classification/assigning taxonomy). mothur has a unique step in which all sequences must be aligned to a template database and any sequences which do not overlap in the same space are removed from the analysis. OTU: Operational taxonomic unit; QC: Quality control.

py. *Assign_taxonomy.py* was used for the classification of each of the representative sequences. The default UCLUST consensus taxonomy assigner and SILVA reference database (Version 111) [22] were used to assign taxonomy.

mothur

The same quality control parameters used in QIIME were used in mothur. Following quality control, the dataset was simplified using the *unique.seqs* command. An alignment was generated using the *align.seqs* command and the SILVA template alignment. The alignment was cleaned using the *screen.seqs* and *filter.seqs* commands. The *pre.cluster* command was run to merge together any reads that were within two base-pair-similarity of a more abundant read.

Reads were given a taxonomic classification using the *classify.seqs* command using the SILVA reference database (Version 111) and the RDP Naïve Bayesian Classifier. Reads that could not be classified at a kingdom level were removed using the *remove.lineages* command and a distance matrix was built using the *dist.seqs* command, discarding distances greater than 0.15. This matrix was then used to generate the OTUs. The *cluster* command to group the reads in OTUs based on the distance matrix; *cluster* utilises the average neighbour algorithm [27]. The average neighbour algorithm works by first creating an OTU between the two most closely related sequences (have the smallest distance). This new OTU then replaces the two sequences in the distance matrix, the distances in the matrix are updated and the process is repeated.

We then obtained consensus taxonomy for each OTU using the *classify.otu* command.

MG-RAST

Post the removal of chimeric reads, reads were uploaded to MG-RAST for sequence analysis under the project ID 10404. The MG-RAST pipeline options that were used in the analysis are as follows: artificial replication reads were removed, reads were screened for host contamination using *H. sapiens* NCBI v36 database as a reference database, and reads were filtered on length (reads greater than 2 standard deviations from median read length were discarded) and ambiguous bases where there was no quality score information available.

Reads were given a taxonomic classification using the 'Best Hit Classification' option in MG-RAST. Best Hit Classification reports the highest scoring annotation(s) for each read. In cases where there are two or more equally high scoring annotations, MG-RAST will report all annotations, this has the effect of inflating the weighting of reads

with an ambiguous taxonomic classification. An example of a multiple annotation is shown in Figure S1.

MG-RAST uses the BLAT algorithm to identify rRNA sequences by searching a reduced RNA database. The UCLUST algorithm is then used to cluster identified rRNA sequences. Again, the representative sequence of each cluster (i.e. the longest read of each cluster) was used to assign taxonomy using the SILVA reference database (Version unknown). A maximum e-value cut-off of 1e-5 and a minimum alignment length of 250 bp were selected. Multiple annotations were identified using the pivot table function in Microsoft Excel 2010 (Microsoft Corporation, Redmond, USA), and were resolved using the RDP Naive Bayesian rRNA Classifier [28,29].

Diversity comparisons and statistical analysis

Diversity was compared at a genus level. Each pipeline can be used to calculate diversity measures; however to maintain consistency in data analysis, we used the Vegan package in R [14] to calculate the genus richness and effective number of genera (as an expression of alpha diversity using the Simpson Index [30]). The overall dissimilarity between the abundance profiles generated by each pipeline was evaluated using the ADONIS and non-metric multidimensional scaling (NMDS) functions in Vegan (using the Bray-Curtis Dissimilarity measure) [14].

Differences between the pipelines were evaluated using ANOVA and the Friedman rank sum test. A significance level of alpha=0.05 was used for all tests. Bonferroni's correction was used for multiple testing corrections where required.

Results

QIIME, mothur and MG-RAST each used a similar workflow. Each pipeline had a default algorithm for most analysis steps (Table 2). While users can choose their preferred algorithm(s) in both QIIME and mothur, this is not possible with MG-RAST. A summary of the functionality and features of the three pipelines is presented in Table 2.

Overall, 159,691 reads from 35 samples (average of 4,563 reads per sample) were used for the comparative analysis of QIIME, mothur and MG-RAST. The number of reads assigned any identity (including an 'unclassified' identity) by each pipeline is shown in Table 3. mothur annotated a slightly higher number of reads (P=0.0547). The number of reads that passed de-multiplexing and quality control in each pipeline is also shown in Table 3.

Analysis of the 35 samples took approximately one hour of computational time in QIIME, approximately ten hours of computational time in mothur and approximately two days of manual data cleaning in MG-RAST to remove multiple annotations of reads.

Taxonomic composition comparisons

The four most abundant phyla Proteobacteria, Actinobacteria, Firmicutes and Bacteroidetes were detected at similar abundances by QIIME, mothur and MG-RAST (Figure 2). MG-RAST left more than 10% of the reads unclassified at the phylum level, significantly more than both QIIME and mothur (P=1.00e-08). Proteobacteria was detected in lower abundance by MG-RAST (30.44%), than by both QIIME and mothur (40.78% and 39.55%, respectively) and mothur detected Actinobacteria at a slightly higher abundance (35.93%) than MG-RAST and QIIME (31.96% and 31.73%, respectively). Verrucomicrobia was detected in very low abundance by both QIIME and mothur (0.15% and 0.03%, respectively), but was not detected by MG-RAST. These differences were not considered statistically significant.

A total of 90 distinct genera were identified across the three pipelines. MG-RAST and mothur were the least similar sharing only 39 genera, while QIIME and mothur were the most similar (sharing 53 genera). QIIME identified the most number of genera (n=70) and MG-RAST identified the least (n=57).

Bifidobacterium was the most abundant genus detected by the three pipelines: QIIME 35.79%, mothur 35.41% and MG-RAST 36.96%. There were notable differences at genus level classifications. While Klebsiella was the second most abundant genus detected by MG-RAST (12.20%), it was detected at very low abundance (<1%) by both QIIME and mothur (P=0.0026). QIIME identified a high abundance of Enterobacter (28.46%), which was identified in very low abundance by MG-RAST (2.59%) and not at all by mothur (P=<0.0001). A summary of the top five genera identified by each pipeline is presented in Table 4 and a complete list of genera identified across the three pipelines (and associated p-values) is presented in Table S1. The taxonomic composition of all samples across each pipeline is available in Table S2.

mothur was unable to classify 28.92% of reads at the genus level. In comparison, QIIME left 10.27% of reads unclassified at the genus level and MG-RAST left 16.46% reads unclassified. This difference was not statistically significant following correction for multiple testing (P=0.0814). The percentage of reads unable to be classified at both the phylum and genus level is shown in Table 3. The majority of reads (83.16%) unable to be classified by mothur at the genus level were from the Enterobacteriaceae family. This was also observed in QIIME, with 63.17% of unclassified reads coming from the Enterobacteriaceae family. The majority of reads unclassified by MG-RAST at the genus level could not be classified to any taxonomic level (70.10%).

Differences in bacteria from the Enterobacteriaceae were the primary difference observed between pipelines. We removed Enterobacteriaceae reads from the read set and compared the pipelines. We found no significant differences between pipelines at the genus level following removal of Enterobacteriaceae reads.

Diversity analysis comparisons

A significant difference was observed in the effective number of genera detected among the three pipelines (P=< 0.0001). A significant difference was also observed among the three pipelines with respect to genus richness (P=<0.0001). Diversity values between the three pipelines are presented in Table 3.

A statistically significant difference was also found among the three pipelines by ADONIS (P=0.01), but the R statistic is low (R=0.0941), suggesting that the effect of the pipeline was negligible. Furthermore, the NMDS plot (using the Bray-Curtis dissimilarity measure) shows that the data did not form distinct clusters based on pipeline (Figure 3).

Discussion

This study compared three 16S rRNA gene analysis pipelines, QIIME, mothur and MG-RAST using a single dataset of human gut microbial read data collected from preterm infants. We found that while little difference exists among the three pipelines with respect to diversity measures and taxonomic classifications, substantial differences exist in usability, particularly with time taken to analyse samples and ease of use.

Composition and diversity comparisons

No statistically significant differences were observed at the phylum

	QIIME	mothur	MG-RAST
License	Open-source	Open-source	Open-source
Implemented in	Python	C++	Perl
Current version (at 13.03.15)	1.9.0	1.34.0	3.5
Cited (according to Scopus at 08.04.15)	1769	2565	722
Website	http://qiime.org/ [34]	http://www.mothur.org/ [35]	http://metagenomics.anl.gov [36]
Web-based interface	YES (http://www.n3phele.com/) Not supported/maintained by the QIIME team	NO	YES (at website above)
Primary usage	Command line	Command line	GUI (at website above)
Amplicon analysis	YES	YES	YES
Whole metagenome shotgun analysis	YES – experimental only	NO	YES
Sequencing technology compatibility	Illumina, 454, Sanger, Ion Torrent, PacBio	Illumina, 454, Sanger, Ion Torrent, PacBio	Illumina, 454, Sanger, Ion Torrent, PacBio
Quality control	YES	YES	YES
16S rRNA gene Databases searched	RDP, SILVA, Greengenes and custom databases	RDP, SILVA, Greengenes and custom databases	M5RNA, RDP, SILVA and Greengenes
Alignment Method	**PyNAST**, MUSCLE, INFERNAL	**Needleman-Wunsch**, blastn, gotoh	**BLAT**
Taxonomic analysis/assignment	**UCLUST**, RDP, BLAST, mothur	Wang/**RDP** approach	**BLAT**
Clustering algorithm	**UCLUST**, CD-HIT, mothur, BLAST	**mothur**, adapts DOTUR and CD-HIT	**UCLUST**
Diversity analysis	alpha and beta	alpha and beta	alpha
Phylogenetic Tree	FastTree	Clearcut algorithm	YES
Chimera detection	**UCHIME**, chimera slayer, BLAST	**UCHIME**, chimera slayer, and more	No
Visualisation	PCA plots, OTU networks, bar plots, heat maps	Dendrograms, heat maps, Venn diagrams, bar plots, PCA plots	PCA plots, heat maps, pie charts, bar plots, Krona and Circos for visualisation
User Support	Forum, tutorials, FAQs, help videos	Forum, SOPs, FAQs, user manual	Video tutorials, FAQs, user manual, 'How to' section on website

Where known, the algorithm used by each pipeline is named. The default algorithm, where known, is bolded. GUI: Graphical User Interface; RDP: Ribosomal Database Project; M5RNA: Non-redundant multisource ribosomal RNA annotation; PyNAST: PythonNAST; MUSCLE: MUltiple Sequence Comparison by Log-Expectation; INFERNAL: INFERence of RNA *Alignment*; BLAST: Basic Local Alignment Search Tool; BLAT: BLAST-Like Alignment Tool; CD-HIT: Cluster Database at High Identity with Tolerance; PCA: Principal Coordinate Analysis; OTU: Operational Taxonomic Unit; FAQ: Frequently Asked Questions; SOPs: Standard Operating Procedures

Table 2: Comparison of the functionality and features of QIIME, mothur and MG-RAST

level, suggesting that the three pipelines provide a comparable overview of sample composition. Differences among the three pipelines were most notable at the genus level, particularly in the classification of members from the Enterobacteriaceae family. The difficulty we experienced in classifying reads from the Enterobacteriaceae family may be a result of the variable regions (V3-V5) of the 16S rRNA gene we targeted. While there is no single variable region in the 16S rRNA gene that can be used to distinguish closely related Enterobacteriaceae bacteria, research by Chakravorty et al. [31] recommends that the combined V3 and V6 regions may be best for distinguishing these bacteria.

In terms of genus richness and effective number of genera, the differences observed among the pipelines are most likely a result of the higher percentage of unclassified reads in mothur as compared to the other pipelines. However, the results from ADONIS and NMDS suggest that there is very little difference in the results generated by the three pipelines.

Usability and functionality comparisons

Each pipeline provides a 'one-stop-shop' for the analysis of 16S rRNA gene sequencing data. However, there are fundamental differences in how each pipeline has been developed. MG-RAST provides an automated service where the user uploads sequencing data to a web application and selects a set of quality control parameters. The data then automatically pass through a series of steps and the user is left to only generate abundance profiles and visualisations. QIIME brings together ('wraps') multiple external programs/algorithms, and enables the user to seamlessly feed output from one program/algorithm into another.

mothur re-implements multiple external programs/algorithms into a single program and has made modifications to several of the re-implemented programs/algorithms to increase speed and improve functionality. This means that the installation of mothur is simpler than the installation of QIIME, because while the mothur pipeline has no external dependencies and is installed as a single program, some of the external dependencies in QIIME must be installed independently of the main pipeline. Both QIIME and mothur require the user to have command line experience; however the documentation and tutorials provided by the teams of both pipelines are comprehensive enough that this is not a hurdle to usability. Analysis with MG-RAST is performed using a graphical user interface (GUI) through a web browser and so it does not need to be installed and requires no programming experience. mothur also has an available GUI for installation, but the command line version is more widely used.

In terms of workflow, the most notable difference is that mothur generates an alignment of the data by aligning query reads with reference 16S rRNA gene sequences in a template alignment database. The alignment is then cleaned to ensure all reads overlap in the same region of the 16S rRNA gene. This process is unique to the mothur pipeline and the inclusion of this step is designed to increase the robustness of assignment of reads into OTUs [32].

	QIIME	mothur	MG-RAST	p-value
Approximate analysis time (for the study dataset)	1 hour	10 hours	2 days	-
Number of reads uploaded	159,691	159,691	159,691	-
Number of reads post demultiplexing and QC	131,661	132,314	131,368	0.695
Number of reads assigned identity	123,909	128,064	123,022	0.0547
Number of unclassified reads at phylum level (%)	104 (0.08)	155 (0.12)	14,199 (11.54)	<0.0001*
Number of unclassified reads at genus level (%)	12,724 (10.27)	37,039 (28.92)	20,253 (16.46)	0.0814
Number of genera identified	60	50	57	-
Genus Richness (median, IQR)	10 (9-15)	8 (5-12)	9 (7-14)	<0.0001*
Effective number of genera (median, IQR)	3 (2-4)	2 (2-3)	3 (3-4)	<0.0001*

*indicates statistically significant difference at alpha=0.05.
QC, quality control
IQR, interquartile range (25-75)

Table 3: Comparison of analysis with QIIME, mothur and MG-RAST.

QIIME	mothur	MG-RAST
Bifidobacterium (35.79%)	*Bifidobacterium* (35.41%)	*Bifidobacterium* (36.96%)
Enterobacter (28.46%)	*Escherichia-Shigella* (12.38%)	*Klebsiella* (12.20%)
Enterococcus (4.09%)	*Enterococcus* (4.27%)	*Escherichia-Shigella* (7.20%)
Clostridium (3.44%)	*Staphylococcus* (3.41%)	*Enterococcus* (4.17%)
Staphylococcus (3.41%)	*Clostridium* (3.26%)	*Veillonella* (3.57%)

NB: abundance of each genus is expressed in parentheses following the genus name *Escherichia* and *Shigella* are difficult to resolve using 16S rRNA gene analysis [31] and as such have been grouped together as one genus.

Table 4: The five most abundant genera detected by QIIME, mothur and MG-RAST.

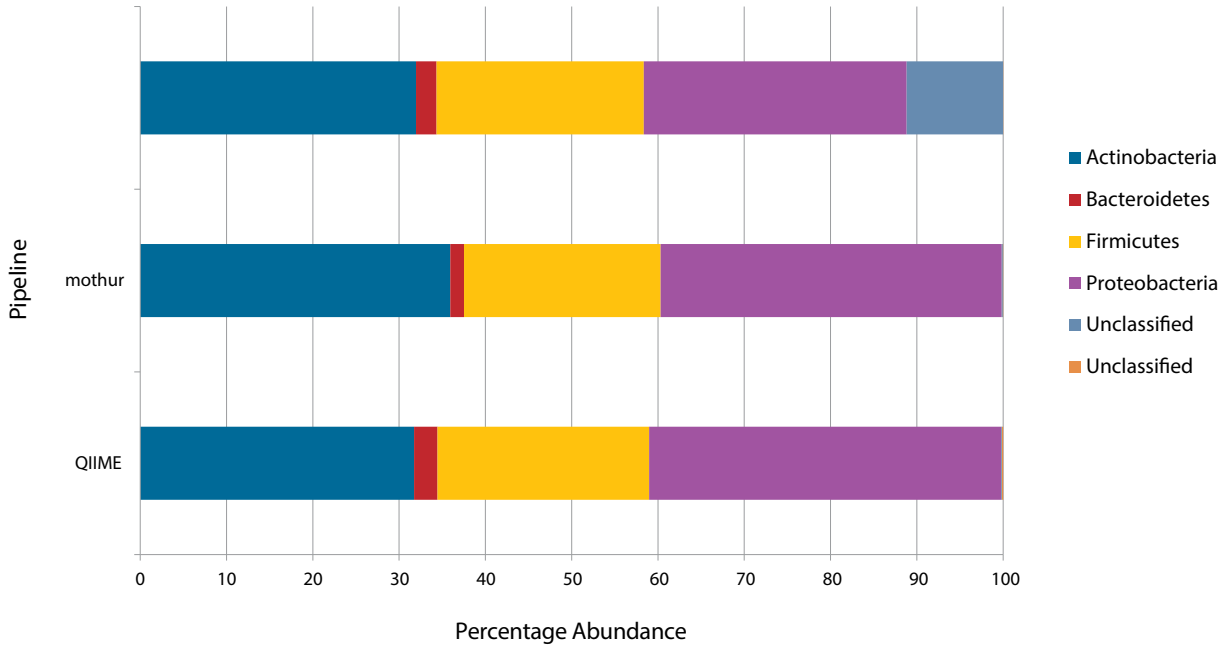

Figure 2: Comparison of taxonomic classification by QIIME, mothur and MG-RAST at the phylum level. This presents the abundance of different phyla detected by QIIME, mothur and MG-RAST. The phyla present in the less than 1% category are: Verrucomicrobia and Fusobacteria.

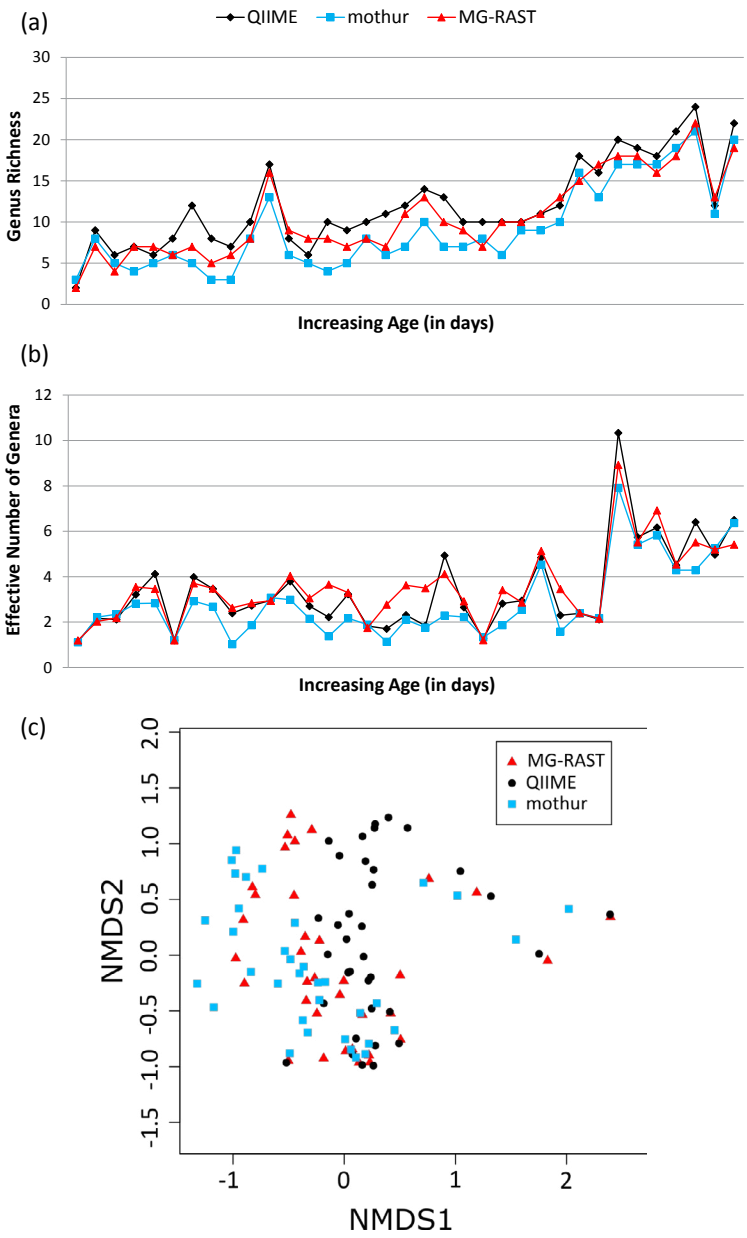

Figure 3: Comparison of diversity measures between QIIME, mothur and MG-RAST. (a) Presents genus richness (i.e. number of identified genera) for each sample (n=35) as determined by the three pipelines, (b) Presents effective number of genera (as a measure of alpha diversity) for each sample (n=35) as determined by the three pipelines. (c) Non-metric multidimensional scaling (NMDS) plot of pair wise Bray-Curtis dissimilarities between all samples processed using the three pipelines. No clear clustering of samples based on pipeline was observed. All diversity measures were calculated using genus level data. For Figure 3(a) and 3(b), samples are on the x-axis and are arranged in order of increasing age.

Analysis with MG-RAST is straightforward. MG-RAST does not require the user to input any commands and it has fewer analysis options than QIIME and mothur, making it more suitable for researchers without bioinformatics or command line experience. The website is easy to navigate and the analysis options are clear and well explained. Once the data have undergone quality control, without post analysis researchers can quickly obtain a descriptive overview of a bacterial population. However, the data produced by MG-RAST require a lot of cleaning due to the multiple annotations of reads. Though it is not difficult to do, cleaning the data is time consuming and would be challenging to complete in a timely manner for large data sets. Because of this, analysis with MG-RAST is the most time consuming. The most common type of multiple annotations observed were where reads were annotated either as both a specific bacterial species and 'unclassified', or annotated as two different species from the same genus (see Figure S1). Furthermore, while analysis can begin immediately in QIIME and mothur, samples must undergo quality control by the MG-RAST team prior to being available for analysis. MG-RAST assigns a priority to data uploaded for analysis based on when the data set will be made

publically available and the wait for private data to undergo quality control can be up to several weeks.

We found QIIME to be more user friendly than mothur. It was easier to understand which command to use in QIIME and it took several attempts to generate a sensible output from mothur, thus our analysis in mothur took a lot longer than in QIIME. We encountered the most difficulty when creating, screening and filtering the alignment in mothur. Initially we screened out too many reads and as a result only 26,477 reads were annotated, 44.61% of which could not be classified to a genus level. We also encountered difficulties when creating the distance matrix, with the operation timing out after 10 hours due to the size of the matrix. The steep learning curve associated with mothur is an important consideration to keep in mind, especially for projects with a short time frame or for research teams with limited informatics experience.

QIIME and mothur are more powerful than MG-RAST, particularly in terms of their statistical capabilities and user freedom. mothur offers more flexibility than QIIME and is likely to be preferred over QIIME and MG-RAST by researchers who are competent at the command line and looking at doing complex amplicon analysis. QIIME is more likely to be preferred for the analysis of a very large datasets, due to its quick analysis time and ease of use. Importantly, the taxonomic summary tables generated by QIIME are the easiest of the three pipelines to adapt to downstream analyses in statistical packages such as R [33].

MG-RAST has some excellent functions. MG-RAST generates a multi-fasta file for each sample when assigning taxonomy. The file contains all reads that were assigned an identity (including unclassified) following the search of the 16S rRNA gene reference database. Each read is identified in the fasta file with a heading shown in Figure S1. This file is easy to access and is extremely useful for resolving the multiple annotations generated by MG-RAST, analysing unclassified reads, and also selecting particular reads to perform downstream analyses such as multiple sequence alignments.

There is no requirement for access to a powerful computer to process multiple samples with MG-RAST, unlike mothur and QIIME, making it easily accessible to all users with an internet connection. Furthermore, MG-RAST acts as a public database for 16S rRNA gene and shotgun metagenomic datasets, which allows comparison and investigations of other publicly available datasets.

There are limitations to this study. We acknowledge that there are some inconsistencies between the analysis methods used by the three pipelines that may impact on comparability. For example, the quality control parameters used in MG-RAST were different to those used in QIIME and mothur, and we were not able to determine the version of the SILVA database used for taxonomic assignment by MG-RAST. This is a proof of principle analysis showing how choice of bioinformatics pipeline can impact the analysis of 16S rRNA gene sequencing data. The strength of this study is that it used a larger dataset than similar comparative analyses [18]. However, given the data used in this study are all of the same sample type and came from the one project, it should be highlighted that different sample types may behave differently in each pipeline according to their true taxonomic composition.

This study used a 454 sequencing dataset, however QIIME, mothur and MG-RAST can be used to analyse data from other sequencing platforms, including Illumina MiSeq; and the information outlined above is relevant to all 16S rRNA gene sequencing datasets, not just those from a 454 sequencing platform. We have used QIIME to process 16S rRNA gene sequencing data from a MiSeq system (data not shown), and found that the advantages of QIIME outlined above also apply to MiSeq data.

QIIME, mothur and MG-RAST are still active years after their initial development and undergo regular updates, highlighting the value of each of these projects to 16S rRNA gene analysis. The differences we observed at the genus level highlight a key limitation of using 16S rRNA gene analysis for genus and species level classification - related bacterial species may have near identical 16S rRNA gene sequences. In fact, even genus identification can be unreliable. For example, QIIME groups together bacteria from the *Escherichia* and *Shigella* genera because they cannot easily be distinguished by their 16S rRNA gene sequence [31]. Additionally, the combination of using only part of the 16S rRNA gene sequence and read errors means that discrimination between species is unlikely. Importantly, this study has shown that QIIME, mothur and MG-RAST are comparable in terms of the phylum they detected and regardless of which pipeline or algorithm is selected you are likely to generate a reliable overview of sample composition.

In the field of bioinformatics there are often a multitude of algorithms, software packages, or pipelines that can be used to perform a single task. Even for experienced bioinformaticians, the choice of which method to use can be confusing. This study provides a comparison and overview of three of the most commonly used bioinformatics pipelines for characterising bacterial communities using the 16S rRNA gene. The results of this study may be used as a resource for people with limited bioinformatics experience when selecting a pipeline to analyse 16S rRNA gene data.

Availability of supporting data

Sequence data has been deposited at MG-RAST (http://metagenomics.anl.gov/) under project number 12656 (Static link= http://metagenomics.anl.gov/linkin.cgi?project=12656).

Competing Interests

The authors declare that there are no competing interests.

Authors' Contributions

Erica Plummer performed the bioinformatics and statistical analysis and drafted the paper. Jimmy Twin carried out the laboratory work. Erica Plummer, Jimmy Twin, Dieter M. Bulach, Suzanne M. Garland and Sepehr N. Tabrizi conceived of the study, and participated in its design and coordination. All authors read, edited and approved the final manuscript.

Acknowledgements

EP was supported by a VLSCI MSc Bioinformatics Bursary from the Victorian Life Sciences Computation Initiative (VLSCI). We would like to thank Mr Simon Gladman for his review of the manuscript.

References

1. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, et al. (2007) The human microbiome project. Nature 449: 804-810.
2. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, et al. (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7: 335-336.
3. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, et al. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75: 7537-7541.
4. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, et al. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9: 386.

5. Riehle K, Coarfa C, Jackson A, Ma J, Tandon A, et al. (2012) The Genboree Microbiome Toolset and the analysis of 16S rRNA microbial sequences. BMC Bioinformatics 13 Suppl 13: S11.

6. Kim OS, Cho YJ, Lee K, Yoon SH, Kim M, et al. (2012) Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. Int J Syst Evol Microbiol 62: 716-721.

7. Soh J, Dong X, Caffrey SM, Voordouw G, Sensen CW (2013) Phoenix 2: a locally installable large-scale 16S rRNA gene sequence analysis pipeline with Web interface. J Biotechnol 167: 393-403.

8. Arndt D, Xia J, Liu Y, Zhou Y, Guo AC, et al. (2012) METAGENassist: a comprehensive web server for comparative metagenomics. Nucleic Acids Res 40: W88-95.

9. Mitra S, Stärk M, Huson DH (2011) Analysis of 16S rRNA environmental sequences using MEGAN. BMC Genomics 12 Suppl 3: S17.

10. Huse SM, Mark Welch DB, Voorhis A, Shipunova A, Morrison HG, et al. (2014) VAMPS: a website for visualization and analysis of microbial population structures. BMC Bioinformatics 15: 41.

11. Stocker G, Snajder R, Rainer J, Trajanoski S, Gorkiewicz G, et al. (Eds) SnoWMAn: High-throughput Phylotyping, Analysis and Comparison of Microbial Communities, In Poster Presentation. The American Society for Microbiology 110th General Meeting; 2010 May 23–27; San Diego California.

12. Angiuoli SV, Matalka M, Gussman A, Galens K, Vangala M, et al. (2011) CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing. BMC Bioinformatics 12: 356.

13. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, et al. (2014) Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res 42: D633-642.

14. Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, et al. (2013) Vegan: ecological diversity - R Package. R package version 2.0-10.

15. Dray S, Dufour AB (2007) The ade4 package: implementing the duality diagram for ecologists. Journal of Statistical Software 22: 1-20.

16. Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20: 289-290.

17. Nilakanta H, Drews KL, Firrell S, Foulkes MA, Jablonski KA1 (2014) A review of software for analyzing molecular sequences. BMC Res Notes 7: 830.

18. D'Argenio V, Casaburi G, Precone V, Salvatore F (2014) Comparative metagenomic analysis of human gut microbiome composition using two different bioinformatic pipelines. BioMed research international 2014: 325340.

19. Garland SM, Tobin JM, Pirotta M, Tabrizi SN, Opie G, et al. (2011) The ProPrems trial: investigating the effects of probiotics on late onset sepsis in very preterm infants. BMC Infect Dis 11: 210.

20. Jacobs SE, Tobin JM, Opie GF, Donath S, Tabrizi SN, et al. (2013) Probiotic effects on late-onset sepsis in very preterm infants: a randomized controlled trial. Pediatrics 132: 1055-1062.

21. Sim K, Cox MJ, Wopereis H, Martin R, Knol J, et al. (2012) Improved detection of bifidobacteria with optimised 16S rRNA-gene based pyrosequencing. PLoS One 7: e32543.

22. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, et al. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41: D590-596.

23. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27: 2194-2200.

24. Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, et al. (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res 21: 494-504.

25. Victorian Life Sciences Computation Initiative.

26. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26: 2460-2461.

27. Legendre P, Legendre L (2012) Cluster analysis. Numerical Ecology. (3rd Edn) Amsertdam: Elsevier.

28. Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73: 5261-5267.

29. Wang Q, Garrity GM, Tiedje JM, Cole JR. (2014)RDP Classifier 2014.

30. Simpson EH (1949) Measurement of diversity. Nature 163.

31. Chakravorty S, Helb D, Burday M, Connell N, Alland D (2007) A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. J Microbiol Methods 69: 330-339.

32. Schloss PD (2013) Secondary structure improves OTU assignments of 16S rRNA gene sequences. ISME J 7: 457-460.

33. 33. R Core Team (2014) R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2014.

34. QIIME Team (2013) QIIME Quantitative Insights Into Microbial Ecology 2013.

35. Schloss P (2008) mothur 2008.

36. MG-RAST Team (2008) mg-RAST metagenomics analysis server 2008.