A Comparison of Two DNA Metagenomic Bioinformatic Pipelines while evaluating the Microbial Diversity in feces of Tanzanian small holder dairy cattle

Felix Kibegwa (mkibegwa@gmail.com)
university of Nairobi, Department of Animal Production
Stomeo Francesca
Biosciences eastern and central Africa - International Livestock Research Institute (BecA-ILRI) Hub
Bett C. Rawlynce
University of Nairobi, Department of Animal Production
Gachuiri K. Charles
University of Nairobi, Department of Animal Production
Mujibi D. Fidalis
USOMI Limited, Suit 13R, Hardy Post, Ushirika road, Karen, Nairobi

Research article

Keywords:

Posted Date: October 11th, 2019

DOI: https://doi.org/10.21203/rs.2.15922/v1

Abstract

Background:

Analysis of shotgun metagenomic data generated from next generation sequencing platforms can be done through a variety of bioinformatic pipelines. These pipelines employ different sets of sophisticated bioinformatics algorithms which may affect the results of this analysis. Furthermore, no conventional assessment technique for estimating the precision of each pipeline exists and few studies have been carried out to compare the characteristics, benefits and disadvantages of each pipeline. In this study we compared two commonly used pipelines for shotgun metagenomic analysis: MetaGenome Rapid Annotation using Subsystem Technology (MG-RAST) and Kraken, in terms of taxonomic classification, diversity analysis and usability using their primarily default parameters

Results:

Overall, the two pipelines detected similar abundance distributions in the three most abundant taxa Proteobacteria, Firmicutes and Bacteroidetes. Within bacterial domain, 497 genera were identified by both pipelines, while an additional 694 and 98 genera were solely identified by Kraken and MG-RAST respectively. 933 species were detected by the two algorithms. Kraken solely detected 3550 species, while MG-RAST identified 557 species uniquely. For archaea, Kraken generated 105 and 236 genera and species respectively while MG-RAST detected 60 genera and 88 species. 54 genera and 72 species were commonly detected by the two methods. Kraken had a quicker analysis time (~4 hours) while MG-RAST took approximately 2 days per sample.

Conclusions:

This study revealed that Kraken and MG-RAST generate comparable results and that a reliable high-level overview of sample is generated irrespective of the pipeline selected. The observed variations at the genus level show that a main restriction is using different databases for classification of the metagenomic data. Specifically, the pipelines could have been limited because some rumen microbes lack reference genomes. The results of this research indicate that a more inclusive and representative classification of microbiomes may be achieved through creation of combined pipelines.

Background

Metagenomics is a high-throughput sequencing (HTS) technique commonly used to investigate complex microbial communities in terms of composition, structure, diversity, and function. This culture independent application has gained importance in microbiological studies over the past decade (Simon & Daniel, 2011) especially in studies of environmental communities (Gilbert at al., 2014; Kopf et al., 2015), industrial quality control processes (Delcenserie et al., 2014), and in understanding the influence of gastro-intestinal microbes on the health of human beings and their well-being (Qin et al., 2010). The phrase metagenomics can be defined by two distinct approaches: targeted and shotgun metagenomics. Targeted metagenomics is also called amplicon-based metagenomics or metagenetics (Esposito & Kirschberg, 2014). This technique focusses exclusively on a genomic marker, that is amplified before sequencing thus greatly reduces the amount of data to be sequenced and analyzed. Shotgun metagenomics on the other hand uses extraction and sequencing of the complete DNA to study the genomic content of a sample. Consequently, this integrated strategy provides a rich image of the microbiota and offers the chance to study the taxonomic classification and functional characteristics of microbial communities simultaneously (Segata et al., 2013). However, shotgun methods are still extremely costly, and the assessment of data is a difficult job not only because of its data size but also due to its complicated data structure (Lindgreen et al., 2016). This is a major impediment to common algorithms.

After the raw reads from metagenomic sequences are generated, the subsequent stage is to evaluate them in order to assess the microbial composition and structure (Siegwald et al., 2017). To achieve this, there are growing numbers of pipelines for bioinformatic assessment (Lindgreen et al., 2016). These tools include CLAssifier based on Reduced K-mers (CLARK) (Ounit et al., 2015); Genomic Origins Through Taxonomic CHAllenge (GOTTCHA) (Freitas et al., 2015); Metagenomics - Rapid Annotation using Subsystems Technology (MG-RAST) – (Meyer et al., 2008); KRAKEN (Wood & Salzberg, 2014); Quantitative Insights Into

Microbial Ecology (QIIME) (Caporaso et al., 2010); Metagenomic Phylogenetic Analysis (MetaPhlAn) (Segata et al., 2012); MOTHUR (Schloss et al., 2009) and metagenomic operational taxonomic units (mOTU) (Sunagawa et al., 2013). These pipelines incorporate several algorithms in order to give the greatest possible analysis options. As a result, they entail extensive bioinformatic knowledge and computational infrastructure which may not be available to users of these analytical procedures. Additionally, individual pipelines propose their own protocols, suggested analytical steps, and reference databases. Thus, without an evaluation protocol, selecting a pipeline thru a specified criterion for a particular function can easily become a daunting job.

Bioinformatics pipelines can be categorized into various groups based on several criteria for example: 1. based on their usage and 2. based on the bioinformatic techniques they use. Based on their usage, these tools can be grouped into: (i) Self-contained analysis pipelines i.e. those that integrate various procedures for quality control, sequence clustering, taxonomy assignment, computing diversity measures and visualizing results and (ii) those that can only be used for a particular step/s in the analysis pipeline (Plummer et al., 2015). Considering the bioinformatic techniques used, the tools can be grouped into: (i) clustering-first approach algorithms e.g. QIIME, MOTHUR, MetaPhlAn mOTU and (ii) assignment-first approach programs e.g. CLARK, GOTTCHA, KRAKEN, MG-RAST. Clustering-first methods, also known as alignment-based methods, begin with an OTU-clustering phase in which sequence reads are collected into OTUs founded on their similarity. A representative sequence is obtained from each cluster and then matched, using a homology search tool, to each sequence of the reference database. Lastly, by checking best alignments, the representative sequence and OTU of which it belongs are allocated to a taxonomic group. The assignment-first approaches, however, first compare all reads to a database, then assign the lowest possible taxonomy to any reads or a lower common ancestor (LCA) for a group of sequences of the same taxonomy within the reference database. Then, based on their annotations, the reads are categorized into distinct taxonomic units (Lindgreen et al., 2016; Siegwald et al., 2017). Of importance is the fact that clustering-first approaches require a high amount of computing resources as such they are almost exclusively the most applied in targeted metagenomics since the data from this approach is greatly reduced. On the other hand, given the complexity of the whole genome shotgun sequence data, assignment-first approaches are recommended since they are aren't resource intensive as the clustering-first approaches

Many previous studies using the available tools for shotgun metagenomics have focused on showing how a single analytical step (e.g., sequence pre-processing, OTU clustering or taxonomic assignment) impacted on the microbial classification in real or simulated datasets (Siegwald et al., 2017). In addition, limited literature evaluates the usability and functions of these tools, which often makes the choice of which technique to use unclear. (Plummer et al., 2015). A study to benchmark the most widely used tools for metagenomic analysis showed that the tools most frequently used were not inherently the most precise, that the most effective tool were not automatically the most time-saving and there was a high level of variation between the available pipelines (Lindgreen et al., 2016). Similarly, a study by D'Argenio et al., (2014) compared the taxonomic and diversity profiles created by MG-RAST and QIIME using Human Gut Microbiome samples. No statistically significant differences in assignments or alpha diversity measures were found in the study; however, there was a significant difference in beta diversity measures between the two pipelines. The researchers also noted that the more accurate assignments were produced by QIIME, primarily due to the high number of reads that MG-RAST could not classify. In contrast, few studies have been undertaken to assess the methodologies available to comprehensively classify the microbiome within the Gastro-Intestinal Tract (GIT) of cattle. This may be attributed to the complexity of the microbial communities that consist of archaea, bacteria, fungi and protozoa (Russell & Rychlik, 2001). For example a previous research by Neves et al., (2017) compared taxonomic compositions of rumen microbial communities using Kraken and an in-house pipeline developed based on Mothur to compare the rumen fluid RNA collected from cattle with different feed conversion ratios (FCR). The study found out that a similar distribution of the most abundant taxa was found in both pipelines at the phylum level; however, unlike Kraken, Mothur was unable to assign sequences to the species level while Kraken's ability to identify microbes was restricted due to an absence of some rumen microbiome reference genomes.

In this study, we used fecal microbial sequence data obtained from thirty six Tanzanian small holder dairy cattle to put forward a comparative analysis of the outcomes of two commonly used assignment-first pipelines, MG-RAST (Meyer et al., 2008) and KRAKEN2 (Wood & Salzberg, 2014), with emphasis on the phylum and genus taxonomic assignments. Functionality and usability of the two pipelines were compared and reviewed. Despite their distinct workflows, these two pipelines were chosen for

evaluation because they are the most frequently used and cited pipelines for analyzing metagenomic data. Additionally, no studies have been carried out to compare these pipelines using shotgun data. This research sheds light on the performance of two commonly used tools and can be of particular use and relevance to scientists who are new to the field or who have limited bioinformatics knowledge to decide which techniques to use in their metagenomics research.

Results

Two bioinformatics methods, Kraken and MG-RAST, were used in this research to acquire taxonomic classifications (bacteria and archaea) of Tanzanian dairy cattle's feces. For most analytical steps, the two tools had a related basic algorithm (Figure 1). However, significant differences in taxonomic assessment, metagenomic function assignment and visualization were noted. Table 1 provides an overview of the characteristics and functionality of the two tools.

MG-RAST and Kraken's taxonomic distribution of microbial profiles

Taking into account the complete amount of microbial species in the specimens, Kraken recognized, in all taxonomic ranks, a greater amount of bacterial and archaeal phylotypes than MG-RAST (Table 2). At the phylum level, bacterial profile findings disclosed a comparable taxa distribution among the most four common species categorized by both pipelines (Tables 2, 3), with Proteobacteria, Firmicutes, Bacteroidetes and Actinobacteria being most abundant and responsible for about 80% of the total microbial population. However, Bacteroidetes was detected in lower abundance by Kraken (9.7%), than by MG-RAST (12.7%) while Actinobacteria had higher abundance (2.9%) in Kraken than MG-RAST (1.25%). Nevertheless, these variations were not regarded as statistically significant. In total both pipelines detected 40 bacterial phyla (Supplementary table 1). Of these phyla, 26 were identified by both pipelines, 12 were solely identified by Kraken while MG-RAST exclusively identified two (Supplementary table 1).

Across the two pipelines, a total of 1289 different genera were detected. Although the two pipelines had some resemblance (497 genera commonly detected), Kraken exclusively identified an extra 694 genera while MG-RAST solely identified 98 genera (Supplementary table 1). The two pipelines identified *Pseudomonas* as the most abundant genus: Kraken 32.64%, and MG-RAST 32.42%. There were significant variations among the most abundant taxa on genus level (relative abundance > 1%) in the two pipelines. Two genera *Comamonas* ($P<0.001$) and *Acinetobacter* ($P = 0.01$) were identified in higher abundance by Kraken while Bacteroides ($P = 0.03$), *Acidovorax* ($P<0.001$) and *Clostridium* ($P<0.001$) had higher abundances in MG-RAST. Table 3 presents an overview of the top genera detected by both pipelines. Kraken and MG-RAST detected 4465 and 1481 species respectively (Table 2). Notable differences in the six most abundant species identified by the two algorithms were the higher abundance of *Prevotella ruminicola* and *Escherichia coli* in Kraken whereas *Pseudomonas fluorescens, Comamonas testosterone, Pseudomonas putida* and *Pseudomonas stutzeri* were more abundant in MG-RAST. A full list of phylotypes (in all taxonomic ranks) recognized across the two pipelines is provided in Supplementary Table 2 (Kraken) and 3 (MG-RAST).

In terms of archaea identification, both methods identified five phyla, with four being identified by both methods. In addition, significant differences were observed in the two methods at the genus and species levels. Kraken generated 105 and 236 genera and species respectively while MG-RAST detected 60 genera and 88 species. 54 genera and 72 species were commonly detected by the two methods (Table 1). Individual algorithm differences in archaea identification can be found in the Supplementary Tables 1, 2 and 3.

Taxa related abundance differences between Lushoto and Rungwe samples

To assess how the two approaches affected biological interpretation of bacteria and archaea composition and community structure, comparisons of fecal microbiota were made between Lushoto and Rungwe samples. Microbial abundance differences between Lushoto and Rungwe data sets were observed to be minimal (< 1% of all microbial population), irrespective of the pipeline (Tables 2 and Supplementary tables 2 and 3). In this regard, within bacteria, one genus *Planococcus* and two

genera *Hafnia* and *Spiroplasma* were detected in Lushoto and Rungwe datasets respectively with MG-RAST classification. This was contrary to Kraken that identified these genera in the two regions. At the species level 22 and 18 species were identified only in Lushoto and Rungwe samples respectively based on Kraken while MG-RAST detected 12 species exclusively in Lushoto and 9 species only in Rungwe. Within archaea, only one difference was observed between the two regions when samples were classified using Kraken, that is *Methanosarcina sp. WH1* was detected in Rungwe only. Assessment of the differences in microbial abundance between Lushoto and Rungwe datasets using an independent t test revealed no significant difference in all microbial taxa detected in both regions. In addition, alpha-diversity indexes (Shannon, Simpson and Chao 1), of bacteria, at the genus level, and archaea, at the species level, indicated no significant difference when they were compared between Lushoto and Rungwe groups within the two pipelines (Figure 2).

Usability and Overall performance

Each pipeline provides avenue for analysis of shotgun metagenomic sequencing data. There are, however, major variations in the development of each pipeline. MG-RAST offers an interactive service where the researcher uploads information to a web application and chooses a number of parameters for quality control. The data then passes through several analytical steps automatically and then the user is left to produce abundance profiles, functional features and visualizations. Moreover, MG-RAST analysis are conducted using a web-based graphical user interface (GUI) making it readily available to all researchers with an internet connection. In addition, to processing multiple samples, it does not need to be installed nor does it require a powerful computer. Furthermore, MG-RAST acts as a public database for metagenomic shotgun datasets and as such investigators can compare and investigate other publicly available datasets. Navigation around the website is easy and the options for analyzing data are clear and well explained. Contrary, analysis with MG-RAST is very time consuming as the outputs require a lot of cleaning because of the multiple read annotations. Although it is not hard to do, data cleaning is time consuming and would be hard to finish in a timely way for big data sets. Besides, whilst in Kraken the analysis can start immediately, the samples must go through a quality control in MG-RAST before they can be analyzed. MG-RAST gives a precedence to data submitted for analysis based on when the data set will be publicly released and the wait for private data to undergo quality control can take up to several weeks.

Kraken on the other hand is a command-based algorithm where the user uses a set of sequential scripts to achieve classification in a custom or default database. The main challenge in Kraken is that this algorithm may be tasking especially if the user has to build their own custom database using genomes found in the Refseq. Because of its quick analysis time Kraken is more likely to be used to analyze a large dataset. Moreover, researchers with command line competence and looking to carry out complex analysis may prefer Kraken due to its increased user freedom. The Kraken pipeline was adapted to include all reference genomes in Refseq, which has led to more bacterial species and phylotypes being identified. However, the findings of the classification of archaea and some of the bacterial species recognized by Kraken should be assessed judiciously as many phylotypes detected have not yet been defined in the rumen. In addition, while Kraken has enhanced taxonomic evaluation at species level, the large amount of unclassified sequences (70%) suggest the need for a clearer taxonomic classification in rumen microbes.

In addition to ease of usage, runtime and memory requirements for shotgun metagenomic algorithms are important factors to consider and should not be underestimated. The run time varied between the two tools with Kraken using 4 hours while MG-RAST took 2 days on average. The runtime of a user with MG-RAST can strongly rely on a number of variables, including present load, software upgrades and work priorities. At the peak memory usage Kraken used 35GB RAM per sample. We were unable to carry out this assessment on MG-RAST as it was only accessible through a website.

Discussion

In this research, the taxonomic results of two metagenomic assessment pipelines, Kraken and MG-RAST, were compared using fecal metagenome data of dairy cattle reared by smallholder farmers in Tanzania. The emergence of high-throughput sequencing has significantly improved our understanding about the ecology and functional ability of different ecosystems

including the GIT of cattle. However, the functional results and the biological interpretation of this information rely heavily on the computational methods used (Siegwald et al., 2017; Simon & Daniel, 2011). We observed that while there is little variation between the two pipelines in terms of taxonomic classifications and diversity measures, there were substantial usability differences, particularly in time taken for analysis of samples and the ease of use.

In this research, both analysis tools showed that the feces of cattle were dominated by Proteobacteria followed by Bacteroidetes and Firmicutes. The dominance of Proteobacteria in these samples, without any health or production effect, is of particular interest given that earlier researchers have found a mechanistic interplay between Proteobacteria, intestinal immune response and inflammation (Maharshak et al., 2013). A recent publication of nursing calves from 5 beef farms with greater concentrations of Proteobacteria comparable to this research by Weese & Jelinski, (2017), proposed that greater concentrations of Proteobacteria could have been a farm-associated effect, possibly from management practices. Specifically, the foregoing authors noted that Proteobacteria-enriched microbiota was observed in farms that had the highest antimicrobial treatment rates leading to their speculation that practices of antimicrobial use could have a wider or cumulative impact on farms in which their recurrent uses result in development, regardless of individual antimicrobial exposures, of a specific microbiota in farm animals. In this study we were unable to confirm this assertion since the sampled farms had no proper recording systems on use of antimicrobials. However, some studies, although in humans, have corroborated this theory (Arboleya et al., 2015; Dardas et al., 2014). Although Bacteroidetes abundance was higher in MG-RAST than Kraken, this difference were not significantly significant. This phylum contains a wide range of individuals who can be found in several ecosystems including; the mammalian and insect guts, soil and both fresh and salt water ecosystems (Ahmed et al., 2018; Gomez-Alvarez et al., 2012; Joynson et al., 2017). A typical characteristic of ecological Bacteroidetes is their capacity to break down complex glycans, for example, agarose, alginate, cellulose, chitin and hemicellulose (Lapébie et al., 2019). It has also been observed that Bacteroidetes are involved in the spread of antimicrobial resistance genes through horizontal gene transfer (Niestępski et al., 2019). Firmicutes were the third most abundant taxonomic group. This phylum is believed to play a crucial part in the harvest of energy. Moreover, members within this phylum can have both positive and negative influences on the host animal. Some species within this phyla, for instance, engage in the degradation of complex organic materials such as cellulose, chitin, xylan lignocellulose and xylose, and even act as useful probiotics and nitrogen fixing agents (Dowd et al., 2008). Conversely several species are potentially hazardous and can cause several diseases in animals and humans (Girija et al., 2013).

Interestingly, both methods had an under-representation of potentially important Cyanobacteria phyla (Supplementary Tables 2 and 3), supporting findings from prior research of low abundance of these phototrophic oxygenic bacteria in dairy and beef cattle rumen (Li & Guan, 2017; Schären et al., 2017). Cyanobacteria can be both heterocystous or non-heterocystous (Nandi & Sengupta, 1998). Although the ruminal environment is commonly deemed anaerobic, significant levels of oxygen in the rumen fluid can be identified (Newbold et al., 1996), suggesting that the occurrence of cyanobacteria in the rumen may be associated with the scavenging of oxygen and the fermentation of sugar under restricted aerobic environments (Neves et al., 2017). Whereas Cyanobacteria has been extensively detected in aqueous and soil environments (Cruz-Martínez et al., 2009; Williams et al., 2004), it is important to point out that the identification of this phylum in the gut of humans raises grave questions regarding their role in aphotic and anaerobic habitats like the rumen (Soo et al., 2014). Recent investigations have shown gut Cyanobacteria to be very conserved but their 16S rRNA phylogenetic genetic tree was different from the photosynthetic ones, this has led to the designation of a new putative class called Melainabacteria (Soo et al., 2014), whose members were able to ferment a variety of sugars in the gut (Di Rienzi et al., 2013). Similar to other previous studies by Girija et al., (2013) and Neves et al., (2017), neither Kraken nor MG-RAST identified Melainabacteria in the samples, showing the need for further research to distinguish their role in the gut of cattle.

The three most dominant genera in both pipelines were *Pseudomonas, Comamonas* and *Acinetobacter*. These genera are not only important in the rumen ecosystem but have also been linked with other environmental roles. For instance, there are many reports indicating that *Pseudomonas* spp. produced antifungal compounds, siderophores and indole acetic acid (IAA). However, they are considered powerful human pathogens that may cause respiratory, urinary and gastrointestinal tract infections (Boricha & Fulekar, 2009). Although low in abundance *Acinetobacter* and species within it are common in nature and some strains are known to be engaged in biodegradation of a variety of pollutants. They are also involved in manufacture of products

such as lipases, proteases, cyanophine, bioemulsifiers and various types of biopolymers (Girija et al., 2013). Furthermore, *Acinetobacter* has been reported to have a role in phosphate solubilization and nitrogen fixation (Desouky, 2003).

In order to fully comprehend the role of the rumen microbiota, it is vital to define organisms at the level of the species since distinct species can have distinct tasks and niches within the same genus. Unlike MG-RAST that used an already preformed database, the Kraken based approach, used a custom reference database assembled based on all identified microbial genomes, at that time. As a result, higher microbiota resolution was generated by Kraken, enabling the program to uniquely identify 3550 species compared to MG-RAST that identified 557 species. However, Kraken is also limited by the lack of all reference genomes for rumen microorganisms. For instance, identification of *Xenorhabdus doucetiae,* a soil bacterium, that had not been earlier recoded in the rumen contents metagenome (Li & Guan, 2017). Identification of this bacterial species may show that Kraken did not correctly identify the microbe, since the reference genome data was based mostly on all microbial genomes annotated in the NCBI database. However, these organisms may have been identified in the rumen, since cattle can eat soil, which makes it possible to detect them temporarily (Neves et al., 2017).

The Archaea domain was dominated by the phylum Euryarchaeota in both pipelines. However, at the genus level *Hyperthermus* and *Methanobrevibacter* were identified as the most predominant genus by Kraken and MG-RAST respectively (Supplementary Tables 2 and 3). Previous studies have reported that *Methanobrevibacter* was the most abundant archaeal population in the rumen based on DNA datasets (Henderson et al., 2015; Kittelmann et al., 2013). However, further studies are needed to determine whether the differences in archaeal abundance between these two pipelines have a methodological influence or are controlled by diet, host animal or management strategies. The differences in the two algorithms were further shown by contrasting results in species identified. For example, Kraken was able to detect *Candidatus Methanoplasma termitum* and *Candidatus Methanomethylophilus alvus,* which were not identified by MG-RAST pipeline. Similar finding was found in a previous study by Neves et al., (2017). These two species encode pathways required for hydrogen-dependent methylotrophic methanogenesis by reduction of methyl substrates, without the ability to oxidize methyl substrates to carbon dioxide (Li et al., 2016). Thus, it is possible that these microbes reside in the rumen. Further, Kraken uniquely identified *Methanogenic archaeon ISO4-H5,* a member of the order Methanomassiliicoccales that had been previously shown to exhibit a genome size of 1.9 Mb and GC content of 54%, similar to *Candidatus Methanoplasma termitum* and *Candidatus Methanomethylophilus alvus (*Li et al., 2016*)*. Given the low relative abundance and the species not being identified by both pipelines, future analysis with databasesenriched with sequences from *Methanogenic archaeon ISO4-H5* as well as its isolation, culture and characterization may provide further evidence of this possibility.

Kraken is by far the quicker pipeline due to the fact that the database is not preloaded into memory by default. Such preloading with a RAMDisk is possible and reduces the execution time of Kraken, but requires RAM space at least equal to the database size. When using the full RefSeq database, this tradeoff should be regarded, which could significantly increase runtime. MG-RAST and other webservers don't require high computing resources, but even so require a decent and steady internet bandwidth, and the users rely on external computer resources that they have no command over.

There are limitations to this study. We recognize that there are some differences between the analytical methods used by the two pipelines that can affect comparability. For instance, the quality control parameters for MG-RAST differs from the quality control parameters used in Kraken, and we were unable to determine the SILVA database version used for MG-RAST taxonomy assignment. This is an evidence of principle study illustrating how bioinformatics pipeline selection can affect metagenomic sequencing data analysis. The strength of this study is that it used a larger dataset. However, given that the data used in this study are all the same type of sample and came from the one project, it should be emphasized that various sample types could be different from the actual taxonomic composition in each pipeline.

Conclusion

There are often many algorithms, software packages, or pipelines in the field of bioinformatics that can be used to conduct a single job. Even for skilled bioinformaticians, it is not simple to choose a single "best tool". This research offers a comparison and overview of two of the most frequently used bioinformatics pipelines used in shotgun metagenomics to characterize

microbial communities. The various performance estimates submitted in this research can be used as a benchmark for scientists when choosing a pipeline to analyze shotgun information based on their particular requirements.

Methods

Fecal Sample Collection, DNA extraction, library construction and sequencing

Fecal samples were collected purposefully from thirty-six (36), adult cross-breed dairy cattle from Lushoto and Rungwe districts in Tanzania. Lushoto district lies between latitudes 4° and 6°S and longitudes 38° to 39°E in Tanga region (Mfune, 2015), while Rungwe district is located in Mbeya region and lies between latitudes 9° 00 and 9° 30 E and longitudes 33 °E and 34°S (Karwani et al., 2016). A clean palpation sleeve and sterile lubricant was used to collect about 250g of individual fecal samples from the rectum of each cattle and a sub-sample transferred into sterile 50 ml falcon tube. Samples were then shipped on ice to the Biosciences east and central Africa (BecA-ILRI) Hub, at the International Livestock Research Institute laboratory where they were stored at –20ºC until further processing.

Fecal DNA was extracted with the QIAamp DNA Stool Mini Kit (Qiagen, USA) according to the manufacturer's instructions using approximately 0.25 g of each fecal sample. Additionally, 2 µl of RNAse A was added during the extraction procedure. The yield and integrity of DNA was determined and quantified using a NanoDrop® ND–2000 UV spectrophotometer (Nano- Drop Technologies, Wilmington, DE) and a Qubit 2.0 fluorimeter (Invitrogen, Carlsbad, CA, United States). Sequencing libraries were prepared using the Nextera XT index kit (Illumina), following the manufacturer's guidelines. The quality of libraries was assessed using the Agilent 2200 TapeStation (Agilent Technologies, Santa Clara, CA, United States) and Qubit 2.0 fluorimeter (Invitrogen, Carlsbad, CA, United States). Finally, the libraries were paired-end (2 × 200 bp) sequenced using an Illumina MiSeq v3 (Illumina) System at the BecA-ILRI Hub.

Bioinformatics and statistical analysis

KRAKEN

Prior to sequence analysis, filters were used to extract low-quality reads from all samples. Visualization of the quality of sequences was done using FastQC software version 0.11.5 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads with an average quality score < 20 were then truncated using FASTX-trimmer a module within the FASTX-toolkit version 0.0.14 (http://hannonlab.cshl.edu/fastx_toolkit/). Following quality control steps, detection of taxa by the kmer approach was done using the Kraken2 (Wood & Salzberg, 2014). In this pipeline we used a custom database, that had been built in the ILRI Research Computing cluster (http://hpc.ilri.cgiar.org/), for Kraken using RefSeq (version 88) complete bacteria (15,947 genomes), and Archaea (311) genome sequences. To build this joint database, the script Kraken-build was used, with default parameters, to set the lowest common ancestors (LCAs). Microbial classification of each pair of sequences was then done on the basis of their annotations at the lowest taxonomic level by Kraken in the customized standard database. In this operation, Kraken's k-mer paths allocated each node a specific weight while improving the sensitivity of the classification of species (Wood & Salzberg, 2014). The Kraken-translate and Kraken-mpa-report scripts provided full taxonomic names associated with each classified sequence and standard ranks for each taxon (Figure 1). The complete Kraken2 database took 4h 2m 9.769s to build on a server with 15 CPUs (2.7 GHz) and 116 GB of RAM, while each sequencing dataset used 35 GB RAM for classification.

MG-RAST

Raw reads were uploaded to MG-RAST for sequence analysis in the metagenomic project (https://www.mg-rast.org/linkin.cgi?project = mgp81260) using the IDs: mgm4754670.3, mgm4754671.3, mgm4755578.3, mgm4755579.3 - mgm4755586.3, mgm4755588.3, mgm4755589.3, mgm4755596.3 - mgm4755600.3 for Rungwe and mgm4755610.3 - mgm4755615.3,

mgm4755949.3 - mgm4755952.3, mgm4756138.3 - mgm4756145.3 for samples from Lushoto. The MG-RAST options used in the study were: removal of artificially replicated reads, screening of sequences for host contamination using B. taurus, UMD v3.0 database, and filtration of low quality sequence reads based on the lowest phred score to be counted as a high-quality base of 15 and to at most this many low phred score bases of 5. Reads were given a taxonomic classification using BLAT (Kent, 2002) and the M5NR database (Wilke et al., 2012). We used default settings where the Minimum percentage Identity Cutoff was set to 60% and the Maximum e-Value Cutoff was at 1×10^5. Reads that did not attain the threshold at the chosen taxonomic level wer3e categorized as “Unclassified”, while sequences not assigned to any taxonomic unit fell in the category called “No Hits”. After taxonomic assignment, MG-RAST created a web page to view, analyze and download results so that they can be used for comparison with other tools (McDonald et al., 2012) (Figure 1).

Statistical analysis

To assess the taxonomic assignment power of the two algorithms, we extracted the outcomes acquired at the phylum, genus and species levels. Paleontological STatistics software package for education and data analysis tool (PAST v3.13), (Hammer et al., 2001) was used to calculate diversity measures. Alpha diversity indices calculated included Chao1 minimal richness index (Chao & Shen, 2003), inverse Simpson diversity index (Hill, 1973; Simpson, 1949) and Shannon diversity index (Shannon & Weaver, 1949). We used a t-test to assess for statistically significant differences between each index and relative abundance of the various taxon assigned by the two tools.

Declarations

Ethics approval and consent to participate

This study was approved and performed following the University of Nairobi’s Faculty of Veterinary Medicine Animal care and use committee (ACUC) guidelines. Animals were handled by experienced animal health professionals to minimize discomfort and injury

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated and/or analyzed during the current study are available in the https://www.mg-rast.org/linkin.cgi?project = mgp81260

Competing interests

The authors declare that they have no competing interests.

Authors’ contributions

FK, RB, FM, GC, SF designed the research project and helped prepare the manuscript. FK, FM conducted the bioinformatics and statistical analysis. FK, FM and SF performed the laboratory analyses. All authors read and approved the final manuscript

Acknowledgements

This work was partly supported by the BecA-ILRI Hub through the Africa Biosciences Challenge Fund (ABCF) program. The ABCF Program is funded by the Australian Department for Foreign Affairs and Trade (DFAT) through the BecA-CSIRO partnership; the Syngenta Foundation for Sustainable Agriculture (SFSA); the Bill & Melinda Gates Foundation (BMGF); the UK Department for International Development (DFID) and the Swedish International Development Cooperation Agency (Sida). We thank the University of Nairobi veterinary farm for providing us with the animals used in this study. Special acknowledgement to Dr. Wellington Ekaya, Valerian Aloo and Eunice Machuka who facilitated the ABCF fellowship and laboratory work at BecA ILRI.

References

Ahmed, V., Verma, M. K., Gupta, S., Mandhan, V., & Chauhan, N. S. (2018). Metagenomic Profiling of Soil Microbes to Mine Salt Stress Tolerance Genes. *Frontiers in Microbiology, 9,* 159. https://doi.org/10.3389/fmicb.2018.00159

Arboleya, S., Sánchez, B., Milani, C., Duranti, S., Solís, G., Fernández, N., … Gueimonde, M. (2015). Intestinal Microbiota Development in Preterm Neonates and Effect of Perinatal Antibiotics. *The Journal of Pediatrics, 166(*3), 538–544. https://doi.org/10.1016/j.jpeds.2014.09.041

Boricha, H., & Fulekar, M. H. (2009). Pseudomonas plecoglossicida as a novel organism for the bioremediation of cypermethrin. *Biology and Medicine, 1(*4), 1–10.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., … Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods, 7(*5), 335–336. https://doi.org/10.1038/nmeth.f.303

Chao, A., & Shen, T.-J. (2003). Nonparametric estimation of Sannon's index of diversity when there are unseen species in ample. *Environmental and Ecological Statistics, 10,* 429–443. https://doi.org/DOI: 10.1023/A:1026096204727

Cruz-Martínez, K., Suttle, K. B., Brodie, E. L., Power, M. E., Andersen, G. L., & Banfield, J. F. (2009). Despite strong seasonal responses, soil microbial consortia are more resilient to long-term changes in rainfall than overlying grassland. *The ISME Journal, 3(*6), 738–744. https://doi.org/10.1038/ismej.2009.16

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. https://doi.org/10.1155/2014/325340

Dardas, M., Gill, S. R., Grier, A., Pryhuber, G. S., Gill, A. L., Lee, Y.-H., & Guillet, R. (2014). The impact of postnatal antibiotics on the preterm intestinal microbiome. *Pediatric Research, 76(*2), 150–158. https://doi.org/10.1038/pr.2014.69

Delcenserie, V., Taminiau, B., Delhalle, L., Nezer, C., Doyen, P., Crevecoeur, S., … Daube, G. (2014). Microbiota characterization of a Belgian protected designation of origin cheese, Herve cheese, using metagenomic analysis. *Journal of Dairy Science, 97(*10), 6046–6056. https://doi.org/10.3168/jds.2014–8225

Desouky, A.-E.-H. (2003). Acinetobacter: environmental and biotechnological applications. *African Journal of Biotechnology, 2(*4), 71–74. https://doi.org/10.5897/AJB2003.000–1014

Di Rienzi, S. C., Sharon, I., Wrighton, K. C., Koren, O., Hug, L. A., Thomas, B.C., … Ley, R. E. (2013). The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria. *ELife, 2,* e01102. https://doi.org/10.7554/eLife.01102

Dowd, S. E., Callaway, T. R., Wolcott, R. D., Sun, Y., McKeehan, T., Hagevoort, R. G., & Edrington, T. S. (2008). Evaluation of the bacterial diversity in the feces of cattle using 16S rDNA bacterial tag-encoded FLX amplicon pyrosequencing (bTEFAP). *BMC Microbiology, 8,* 125. https://doi.org/10.1186/1471–2180–8–125

Esposito, A., & Kirschberg, M. (2014). How many *16S*-based studies should be included in a metagenomic conference? It may be a matter of etymology. *FEMS Microbiology Letters, 351(*2), 145–146. https://doi.org/10.1111/1574–6968.12375

Freitas, T. A. K., Li, P.-E., Scholz, M. B., & Chain, P. S. G. (2015). Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. *Nucleic Acids Research, 43(*10), e69–e69. https://doi.org/10.1093/nar/gkv180

Gilbert, J. A., Jansson, J. K., & Knight, R. (2014). The Earth Microbiome project: successes and aspirations. *BMC Biology, 12(*1), 69. https://doi.org/10.1186/s12915–014–0069–1

Girija, D., Deepa, K., Xavier, F., Antony, I., & Shidhi, P. R. (2013). Analysis of cow dung microbiota-A metagenomic approach. *Indian Journal of Biotechnology, 12(*3), 372–378. https://doi.org/Analysis.

Gomez-Alvarez, V., Revetta, R. P., & Santo Domingo, J. W. (2012). Metagenomic analyses of drinking water receiving different disinfection treatments. *Applied and Environmental Microbiology, 78(*17), 6095–6102. https://doi.org/10.1128/AEM.01018–12

Hammer, O., Harper, D. A. T., & Ryan, P. D. (2001). PAST: paleontological statistics software package for education and data analysis, ver. 1.89. *Palaeontol Electron, 2,* 1–31. Retrieved from http://www.researchgate.net/publication/228393561_PASTPalaeontological_statistics_ver._1.89/file/32bfe5135d45cd6b3b.pdf

Henderson, G., Cox, F., Ganesh, S., Jonker, A., Young, W., & Janssen, P. H. (2015). Rumen microbial community composition varies with diet and host, but a core microbiome is found across a wide geographical range. *Scientific Reports, 5(*1), 14567. https://doi.org/10.1038/srep14567

Hill, M. O. (1973). Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology, 54(*2), 427–432. https://doi.org/10.2307/1934352

Joynson, R., Pritchard, L., Osemwekha, E., & Ferry, N. (2017). Metagenomic Analysis of the Gut Microbiome of the Common Black Slug Arion ater in Search of Novel Lignocellulose Degrading Enzymes. *Frontiers in Microbiology, 8,* 2181. https://doi.org/10.3389/fmicb.2017.02181

Karwani, G., Lulandala, L., Kimaro, A., & Msigwa, Z. (2016). The role of short rotation coppice technology in fuelwood supply in Rungwe district, Tanzania. *International Journal of Agricultural Research, Innovation and Technology, 6(*1), 41–46. https://doi.org/10.3329/ijarit.v6i1.29211

Kent, W. J. (2002). BLAT—the BLAST-like alignment tool. *Genome Research, 12(*4), 656–664. https://doi.org/10.1101/gr.229202

Kittelmann, S., Seedorf, H., Walters, W. A., Clemente, J. C., Knight, R., Gordon, J. I., & Janssen, P. H. (2013). Simultaneous Amplicon Sequencing to Explore Co-Occurrence Patterns of Bacterial, Archaeal and Eukaryotic Microorganisms in Rumen Microbial Communities. *PLoS ONE, 8(*2), e47879. https://doi.org/10.1371/journal.pone.0047879

Kopf, A., Bicak, M., Kottmann, R., Schnetzer, J., Kostadinov, I., Lehmann, K., … Glöckner, F. O. (2015). The ocean sampling day consortium. *GigaScience, 4(*1), 27. https://doi.org/10.1186/s13742–015–0066–5

Lapébie, P., Lombard, V., Drula, E., Terrapon, N., & Henrissat, B. (2019). Bacteroidetes use thousands of enzyme combinations to break down glycans. *Nature Communications, 10(*1), 2043. https://doi.org/10.1038/s41467–019–10068–5

Li, F., & Guan, L. L. (2017). Metatranscriptomic Profiling Reveals Linkages between the Active Rumen Microbiome and Feed Efficiency in Beef Cattle. *Applied and Environmental Microbiology, 83(*9). https://doi.org/10.1128/AEM.00061–17

Li, F., Henderson, G., Sun, X., Cox, F., Janssen, P. H., & Guan, L. L. (2016). Taxonomic Assessment of Rumen Microbiota Using Total RNA and Targeted Amplicon Sequencing Approaches. *Frontiers in Microbiology, 7,* 987. https://doi.org/10.3389/fmicb.2016.00987

Li, Y., Leahy, S. C., Jeyanathan, J., Henderson, G., Cox, F., Altermann, E., … Attwood, G. T. (2016). The complete genome sequence of the methanogenic archaeon ISO4-H5 provides insights into the methylotrophic lifestyle of a ruminal representative of the Methanomassiliicoccales. *Standards in Genomic Sciences, 11(*1), 59. https://doi.org/10.1186/s40793–016–0183–5

Lindgreen, S., Adair, K. L., & Gardner, P. P. (2016). An evaluation of the accuracy and speed of metagenome analysis tools. *Scientific Reports, 6(*1), 19233. https://doi.org/10.1038/srep19233

Maharshak, N., Packey, C. D., Ellermann, M., Manick, S., Siddle, J. P., Huh, E. Y., … Carroll, I. M. (2013). Altered enteric microbiota ecology in interleukin 10-deficient mice during development and progression of intestinal inflammation. *Gut Microbes, 4(*4), 316–324. https://doi.org/10.4161/gmic.25486

McDonald, D., Clemente, J. C., Kuczynski, J., Rideout, J. R., Stombaugh, J., Wendel, D., … Caporaso, J. G. (2012). The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *GigaScience, 1(*1), 7. https://doi.org/10.1186/2047–217X–1–7

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., … Edwards, R. A. (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics, 9(*1), 386. https://doi.org/10.1186/1471–2105–9–386

Mfune, R. L. (2015). *Epidemiological study of bovine brucellosis in smallholder dairy cattle in Lushoto and Rungwe districts, Tanzania* (Sokoine University of Agriculture). Retrieved from http://www.suaire.suanet.ac.tz:8080/xmlui/handle/123456789/1576

Nandi, R., & Sengupta, S. (1998). Microbial Production of Hydrogen: An Overview. *Critical Reviews in Microbiology, 24(*1), 61–84. https://doi.org/10.1080/10408419891294181

Neves, A. L. A., Li, F., Ghoshal, B., McAllister, T., & Guan, L. L. (2017). Enhancing the Resolution of Rumen Microbial Classification from Metatranscriptomic Data Using Kraken and Mothur. *Frontiers in Microbiology, 8,* 2445. https://doi.org/10.3389/fmicb.2017.02445

Newbold, C. J., Wallace, R. J., & McIntosh, F. M. (1996). Mode of action of the yeast Saccharomyces cerevisiae as a feed additive for ruminants. *The British Journal of Nutrition, 76(*2), 249–261. https://doi.org/10.1079/bjn19960029

Niestępski, S., Harnisz, M., Korzeniewska, E., Aguilera-Arreola, M. G., Contreras-Rodríguez, A., Filipkowska, Z., & Osińska, A. (2019). The emergence of antimicrobial resistance in environmental strains of the Bacteroides fragilis group. *Environment International, 124,* 408–419. https://doi.org/10.1016/J.ENVINT.2018.12.056

Ounit, R., Wanamaker, S., Close, T. J., & Lonardi, S. (2015). CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics, 16(*1), 236. https://doi.org/10.1186/s12864–015–1419–2

Plummer, E., Twin, J., Bulach, D. M., Garl, S. M., & Tabrizi, S. N. (2015). A Comparison of Three Bioinformatics Pipelines for the Analysis of Preterm Gut Microbiota using 16S rRNA Gene Sequencing Data. *Journal of Proteomics & Bioinformatics, 8(*12), 1–9. https://doi.org/10.4172/jpb.1000381

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., … Wang, J. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature, 464(*7285), 59–65. https://doi.org/10.1038/nature08821

Russell, J. B., & Rychlik, J. L. (2001). Factors that alter rumen microbial ecology. *Science (New York, N. Y.), 292(*5519), 1119–1122. https://doi.org/10.1126/SCIENCE.1058830

Schären, M., Drong, C., Kiri, K., Riede, S., Gardener, M., Meyer, U., … Dänicke, S. (2017). Differential effects of monensin and a blend of essential oils on rumen microbiota composition of transition dairy cows. *Journal of Dairy Science, 100(*4), 2765–2783. https://doi.org/10.3168/jds.2016–11994

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., … Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology, 75(*23), 7537–7541. https://doi.org/10.1128/AEM.01541–09

Segata, N., Boernigen, D., Tickle, T. L., Morgan, X. C., Garrett, W. S., & Huttenhower, C. (2013). Computational meta'omics for microbial community studies. *Molecular Systems Biology, 9(*1), 666. https://doi.org/10.1038/msb.2013.22

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., & Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods, 9(*8), 811–814. https://doi.org/10.1038/nmeth.2066

Shannon, C. E., & Weaver, W. (1949). The Mathematical Theory of Communication. *The Mathematical Theory of Communication, 27(*4), 117. https://doi.org/10.2307/3611062

Siegwald, L., Touzet, H., Lemoine, Y., Hot, D., Audebert, C., & Caboche, S. (2017). Assessment of Common and Emerging Bioinformatics Pipelines for Targeted Metagenomics. *PLOS ONE, 12(*1), e0169563. https://doi.org/10.1371/journal.pone.0169563

Simon, C., & Daniel, R. (2011). Metagenomic analyses: past and future trends. *Applied and Environmental Microbiology, 77(*4), 1153–1161. https://doi.org/10.1128/AEM.02345–10

Simpson, E. H. (1949). Measurement of Diversity. *Nature, 163(*4148), 688–688. https://doi.org/10.1038/163688a0

Soo, R. M., Skennerton, C. T., Sekiguchi, Y., Imelfort, M., Paech, S. J., Dennis, P. G., … Hugenholtz, P. (2014). An Expanded Genomic Representation of the Phylum Cyanobacteria. *Genome Biology and Evolution, 6(*5), 1031–1045. https://doi.org/10.1093/gbe/evu073

Sunagawa, S., Mende, D. R., Zeller, G., Izquierdo-Carrasco, F., Berger, S. A., Kultima, J. R., … Bork, P. (2013). Metagenomic species profiling using universal phylogenetic marker genes. *Nature Methods, 10(*12), 1196–1199. https://doi.org/10.1038/nmeth.2693

Weese, J. S., & Jelinski, M. (2017). Assessment of the Fecal Microbiota in Beef Calves. *Journal of Veterinary Internal Medicine, 31(*1), 176–185. https://doi.org/10.1111/jvim.14611

Wilke, A., Harrison, T., Wilkening, J., Field, D., Glass, E. M., Kyrpides, N., … Meyer, F. (2012). The M5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. *BMC Bioinformatics, 13(*1), 141. https://doi.org/10.1186/1471–2105–13–141

Williams, M. M., Domingo, J. W. S., Meckes, M. C., Kelty, C. A., & Rochon, H. S. (2004). Phylogenetic diversity of drinking water bacteria in a distribution system simulator. *Journal of Applied Microbiology, 96(*5), 954–964. https://doi.org/10.1111/j.1365–2672.2004.02229.x

Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology, 15(*3), R46. https://doi.org/10.1186/gb–2014–15–3-r46

Tables

Table 1: Comparison of the functionality and features of MG-RAST and KRAKEN. "(E)" Indicates if the tool infers Eukaryotic taxa and/or functional analysis. GUI means Graphical User Interface, and "(R)" The server recognizes paired end data but seems to treat reads separately. Part of this figure was adapted from the pipeline published by (Plummer et al., 2015) .

	KRAKEN	MG-RAST
License	Open-source	Open-source
Implemented in	C++ and Perl	Perl
Current version (at 23.05.19)	v2.0.8-beta	4.0.3
Website	http://ccb.jhu.edu/software/kraken/ (Wood & Salzberg, 2014)	https://www.mg-rast.org/ (Meyer et al., 2008)
Web-based interface	NO	YES (at website above)
Primary usage	Command line	GUI (at website above)
Sequencing technology compatibility	Illumina, 454, Sanger, Ion Torrent, PacBio	Illumina, 454, Sanger, Ion Torrent, PacBio
Quality control	NO	YES
Taxonomic analysis/assignment	k-mers	BLAT
Taxonomy	Yes	Yes (E)
Function	No	Yes
Fastq	Yes	Yes
Zipped	Yes	Yes
Paired	Yes	Yes (R)
Diversity analysis	NO	alpha
Phylogenetic Tree	NO	YES
Visualization	NO	PCA plots, heat maps, pie charts, bar plots, Krona and Circos for visualisation

Table 2: Evaluation of taxonomic phylotypes by each technique.

Phylotypes	Kraken		MG-RAST		Commonly detected phylotypes
	Lushoto (No)	Rungwe (No)	Lushoto (No)	Rungwe (No)	
Bacteria					
Phyla	38	38	28	28	26
Genera	1191	1191	595	596	497
Species	4462	4465	1479	1481	933
Archaea					
Phyla	5	5	5	5	4
Genera	105	105	60	60	54
Species	235	236	88	88	72

Table 3: Most abundant bacteria according to the two classification approaches

Taxa	Kraken		MG-RAST	P Value
	Mean ± SE (%)		Mean ± SE (%)	
Phyla				
Proteobacteria	75.92 ± 4.1		75.12 ± 3.06	0.88
Firmicutes	9.69 ± 2.2		9.29 ± 1.63	0.88
Bacteroidetes	9.22 ± 1.41		12.7 ± 1.54	0.1
Actinobacteria	2.9 ± 0.41		1.25 ± 0.13	<0.001
Tenericutes	0.72 ± 0.11		0.19 ± 0.05	<0.001
Genus				
Pseudomonas	32.64 ± 4.1	32.42 ± 3.29	0.97	
Comamonas	8.38 ± 1.34	3.57 ± 0.54	<0.001	
Acinetobacter	6.72 ± 1.82	1.87 ± 0.56	0.01	
Janthinobacterium	2.2 ± 1.36		1.08 ± 0.22	0.42
Bacteroides	1.83 ± 0.4		3.72 ± 0.75	0.03
Acidovorax	1.61 ± 0.23		6.18 ± 0.83	<0.001
Stenotrophomonas	1.85 ± 0.35		1.75 ± 0.3	0.83
Clostridium	1.19 ± 0.26		2.97 ± 0.54	<0.001
Species				
Pseudomonas fluorescens	6.77 ± 1.69		21.67 ± 2.81	<0.001
Comamonas testosteroni	1.03 ± 0.35		3.57 ± 0.54	<0.001
Pseudomonas putida	1.34 ± 0.38		2.34 ± 0.18	0.02
Pseudomonas stutzeri	1.46 ± 0.49		1.66 ± 0.37	0.74
Escherichia coli	1.22 ± 0.21		0.45 ± 0.13	<0.001
Prevotella ruminicola	0.83 ± 0.13		0.21 ± 0.05	<0.001

Table 4: Most abundant archaea according to the two classification approaches

Taxa	Kraken	MG-RAST	P Value
	Mean ± SE (%)	Mean ± SE (%)	
Phyla			
Euryarchaeota	90.96 ± 0.41	94.44 ± 0.44	<0.001
Crenarchaeota	6.59 ± 0.27	4.45 ± 0.39	<0.001
Thaumarchaeota	2.3 ± 0.25	0.54 ± 0.06	<0.001
Korarchaeota	0.07 ± 0.03	0.49 ± 0.08	<0.001
Genus			
Methanocaldococcus	8.17 ± 1.47	2.4 ± 0.16	<0.001
Methanosarcina	7.62 ± 0.51	12.83 ± 0.82	<0.001
Thermococcus	6.79 ± 0.51	2.51 ± 0.19	<0.001
Methanocorpusculum	3.53 ± 0.59	6.39 ± 0.98	0.01
Species			
Methanobrevibacter ruminantium	3.62 ± 0.7	9.51 ± 1.28	<0.001
Methanococcus maripaludis	2.82 ± 0.27	3.42 ± 0.23	0.1
Methanobrevibacter smithii	1.84 ± 0.27	15.87 ± 1.85	<0.001
Methanocorpusculum labreanum	1.97 ± 0.54	6.39 ± 0.98	<0.001
Methanosarcina barkeri	1.6 ± 0.19	3.96 ± 0.23	<0.001

Figures

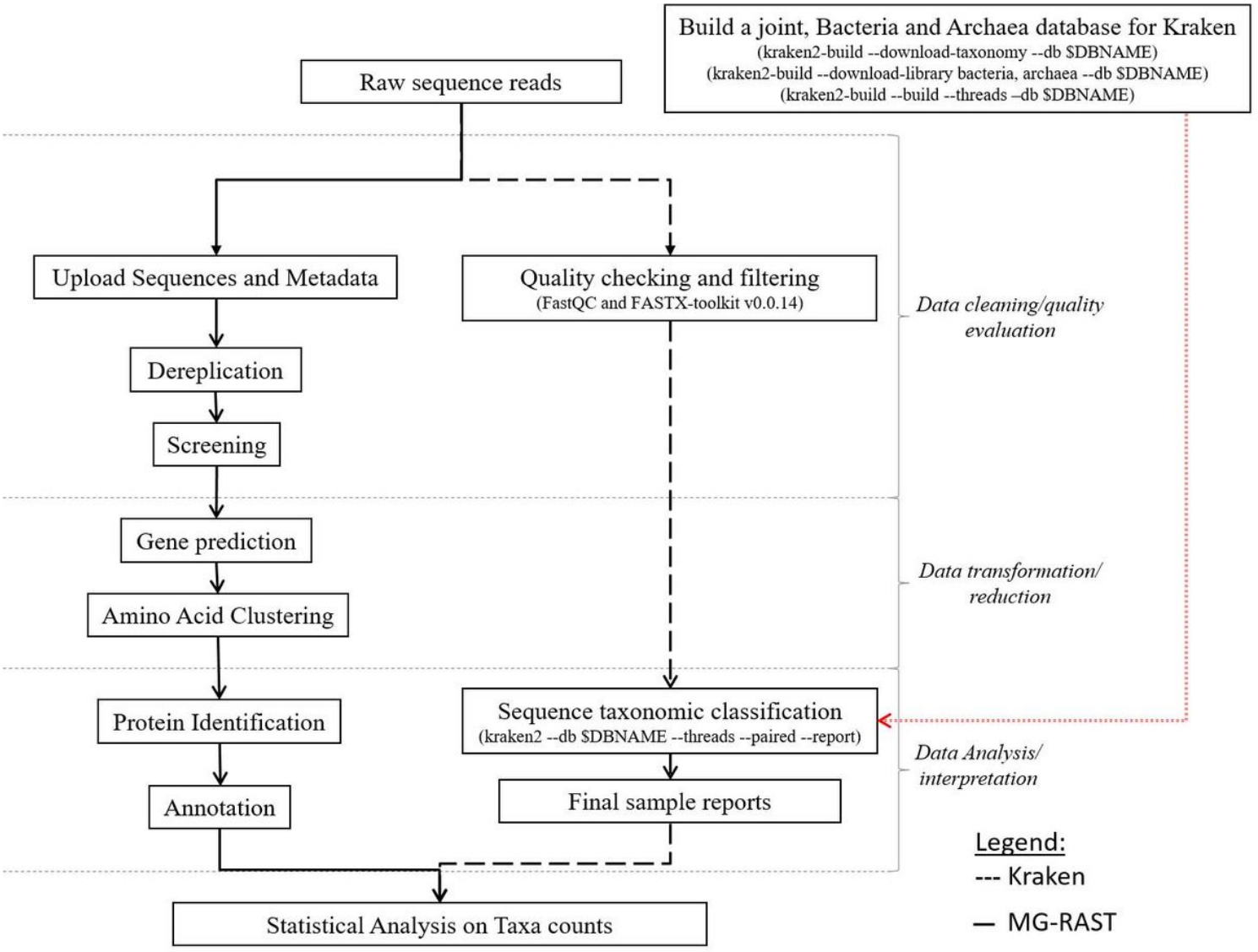

Figure 1

Overview of the workflow used (MG-RAST and Kraken) presenting software parameters used to analyze the data. MG-RAST has two an additional step for data transformation and reduction.

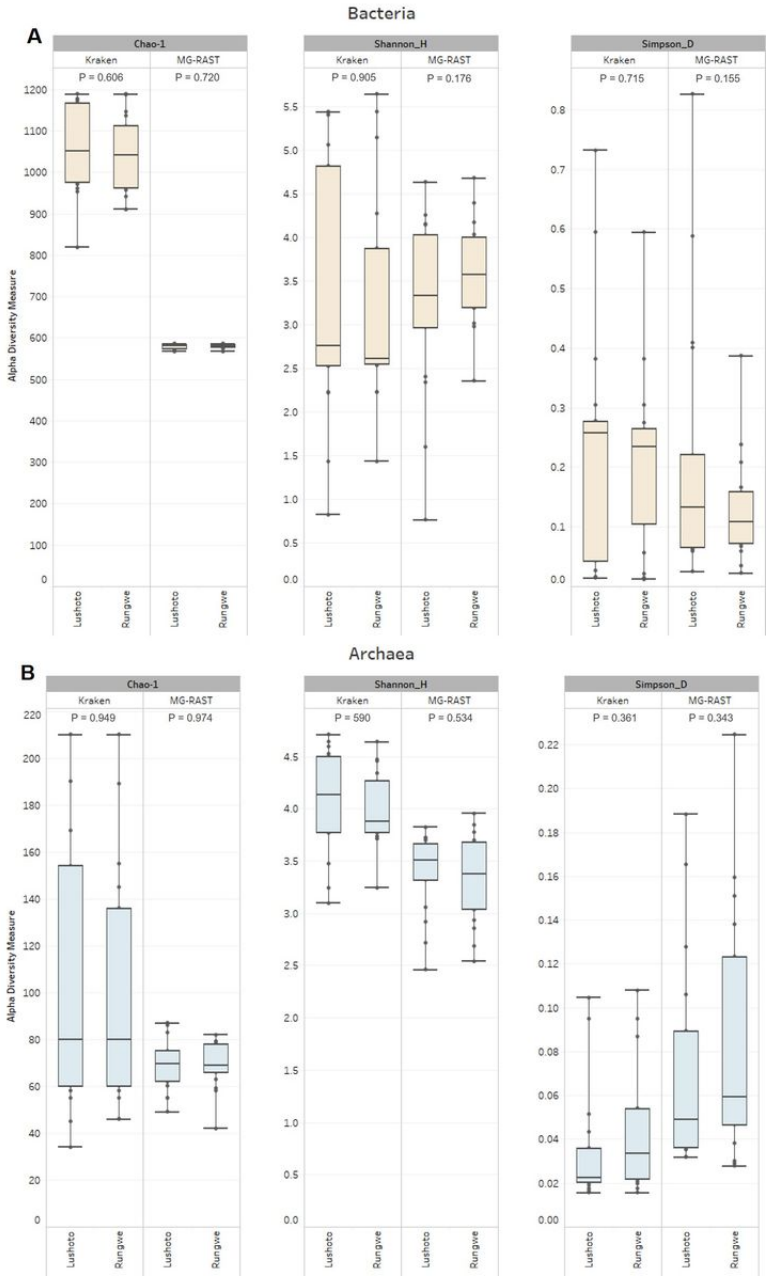

Figure 2

Alpha diversity matrices of bacteria (genus level), A, and archaea (species level), B, between Lushoto and Rungwe samples

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTable2.xls
- SupplementaryTable3.xls
- SupplementaryTable1.xls