Metagenomics : beyond the horizon of current implementations and methods

Khachatryan, L.

Chapter 2

Taxonomic classification and abundance estimation using 16S and WGS - a comparison using controlled reference samples

L. Khachatryan[1], R. H. de Leeuw[1], M. E. M. Kraakman[2], N. Pappas[3], M. te Raa[1], H. Mei[3], P. de Knijff[1], and J. F. J. Laros[1,4]

1 Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands
2 Department of Microbiology, Leiden University Medical Center, Leiden, The Netherlands
3 Sequencing Analysis Support Core, Leiden University Medical Center, Leiden, The Netherlands
4 Clinical Genetics, Leiden University Medical Center, Leiden, The Netherlands

Forensic Science International: Genetics, 2020 46:102275 doi 10.1016/j.fsigen.2020.102257

2.1 Background

In recent years, metagenomics - the genomic analysis of microorganisms by direct extraction of DNA from an environmental sample - has become the most rapidly developing branch of microbiology [161, 162, 163]. The interest in metagenomics has grown drastically due to the expanding number of studies showing that the vast majority of microorganisms cannot be grown under laboratory conditions [35, 164, 165, 166]. The possibility of culture-free investigation of microbial biodiversity directly from an environmental habitat led to many amount of studies benefiting a wide range of fields such as human health [167, 168, 169, 170, 171], ecology [172, 173], agriculture [174, 175, 176], forensics [177, 178], food and drugs production [54, 55, 179]. Taxonomic profiling of metagenomic data is the key step during the data analysis, allowing researchers to understand the structure of a microbiome and to estimate the abundances of the organisms living in it. The main goal of this study is to compare different data types and methods for taxonomic profiling of metagenomic data sets with known abundance distributions of inhabitants.

The most common technique to investigate microbiome composition is amplicon-based sequencing of the 16S rRNA gene [180, 181]. This relatively short (~1500 bp) gene is universal among bacteria and archaea [70, 71]. There are in total nine hypervariable regions in the 16S rRNA gene that provide phylogenetic signatures on different taxonomic levels. Hypervariable regions are surrounded by highly conserved sequences, which are used for primer design. The analysis of 16S metagenomic datasets is usually performed in combination with one of several curated databases that contain annotated sequences of the 16S rRNA gene or its parts [182]. The most commonly used 16S-specific databases are RDB [74, 75], GreenGenes [77] and SILVA [76]. Analysis of 16S data is now routine for metagenomic-associated projects, though many studies demonstrated a number of biases associated with this type of data that make the validity of this approach questionable. Several reports stressed uneven coverage of microorganisms' diversity spectrum by common PCR primers for the 16S rRNA gene amplification [183, 184, 185, 186, 83, 84]. Second, the 16S rRNA gene does not have a correct phylogenetic relationship within particular taxa [81, 82]. The fact that bacteria and archaea might carry different copy numbers of the 16S rRNA gene in their genomes seriously influences a reliable abundance estimation after analysis of 16S data [187]. Additionally, the choice of a specific hypervariable region and the reference database for the subsequent analysis requires *a priori* knowledge about the investigated metagenome. Lastly, 16S data cannot be used to investigate the metagenome functional profile, nor does it provide any information about eukaryotic or viral members of the microbial community. The applicability of 16S data was shown for a set of forensic studies. For example, 16S data was successfully used for body fluid recognition [188] or matching between individuals' skin datasets and touched objects [62, 63]. The success of such analyses,

however, does not imply that a 16S-based analysis of all metagenomic data is reliable (or possible).

Apart from 16S, there are other methods that use rRNA genes to investigate microbial diversity. Among them are 23S, 5S, 12S and various combinations [189, 190, 191]. Other methods like the IS-pro approach use 16S-23S ribosomal interspace fragment lengths to analyse microbial communities [192]. Although these methods are very suitable for some specific tasks, they are not as widely applied as 16S. Several recent studies are based on targeting other genes in addition to 16S in order to determine the cell type of the forensic traces [193] or to perform skin sample identification using only microbial targeting genes [59, 194]. These studies also suggest that traditional 16S data is not always sufficient for a meaningful metagenomic analysis of forensic traces.

In recent years, the number of metagenomic studies based on the whole genome shotgun (WGS) sequencing data type has grown [90, 195, 196, 197, 198]. Among the main reasons for this are advantages in sequencing techniques allowing for the generation of sufficient number of high-quality reads for the WGS datasets, and bioinformatics algorithms to perform subsequent analysis of the big data. Though using WGS data avoids the biases introduced by 16S, it requires more computationally intense analysis, as well as higher sequencing costs.

While many studies in the field of forensics are based on the analysis of 16S data [199], "the capacity of WGS data of microbiomes to aid in forensic investigations by connecting objects and environments to individuals has been poorly investigated" [200]. Presently, WGS experiments are reserved for those studies for which analysis beyond the taxonomical assignment is required: investigating the microbiomes' functional profile, correlation between metagenome and host genome, search for the possible virulent genes, etc. The vast majority of taxonomical annotations is still performed by using only 16S data, despite all known disadvantages of the method [90]. One of the reasons for that is the lack of a well-performed benchmark study, comparing 16S and WGS data types. The vast majority of existing metagenomics benchmarks are created in order to evaluate the accuracy of various metagenomic profiles and comprise either only 16S [201] or only WGSdata [202, 203, 204, 205, 206, 117, 207, 208]. Existing benchmarks that can be used to compare 16S and WGS data types are *in-silico* created and based on a random set of bacterial species, lacking the information about whether or not the selected set of organisms might live together in the same environment [97]. One of the main goals of this study is the creation of a set of benchmarks allowing to compare the 16S and WGS data types using a set of in-vitro DNA mixes of bacteria species inhabiting skin.

Over the last decade, the number of different techniques for metagenomics data analysis has grown remarkably. The tools used for performing the taxonomical annotation, can be split into several groups based on the following criteria: strategy for reads assignment (alignment or matching based on the *k*-mers or sequences signatures); the database against which the search is performed; the proportion of

reads participating in the profiling (all reads, only one read per read group, only reads with particular features).

To investigate which type of metagenomic data is preferable for accurate taxonomic annotation, as well as to test which method of reads assignment yields more precise output, we created a series of bacterial mixes with known content. Each metagenomic mix incorporated 14 to 15 bacterial species belonging to 7 distinct bacterial genera. Each mix had a distinct distribution of the species abundances. For the analysis we selected two popular tools: Centrifuge [111] and MG-RAST [118]. These allow analysis of amplicon and WGS sequencing data and both perform the metagenome profiling by a comparison of sequencing data to a reference database. However, the strategies for metagenome profiling they exploit are different.

We did not include other popular tools for metagenomic analysis in this study as they either have a similar analysis strategy as the tools described above or are designed only for WGS or amplicon data analysis. In many studies, QIIME [100], objectively the most popular tool for amplicon data analysis, was shown to perform with the same accuracy as the MG-RAST pipeline for 16S rRNA sequencing data [209].

2.2 Materials and Methods

2.2.1 DNA extraction and concentration measurement

Laboratory pure cultures of 15 bacterial species that frequently inhabit human skin (Table 2.2) were grown with gentle shaking overnight at 37°C. Genomic DNA was isolated with the Easy-DNA™ gDNA Purification Kit (Invitrogen™ Thermo Fisher Scientific) using the standard protocol with ethanol precipitation [210]. RNA contamination was removed using RNase A (Roche) and the DNA was stored at 4°C. DNA concentrations were measured with the Qubit 3.1 Fluorometer (Invitrogen™).

2.2.2 Metagenomic mixes creation

Four bacterial mixes with known genome abundances were created for this research. In order to achieve the desired species abundances, the estimated genome size and the measured DNA concentration for each bacteria were used. One mix was created to have a uniform- and other three mixes an exponential ($\lambda = 1/6$, $\lambda = 1/2$ and $\lambda = 5/6$) distribution of species abundances. From here on, these mixes are referred to as EQ, EXP16, EXP12 and EXP56 respectively. Due to technical reasons, *Corynebacterium jeikeium* was included only in EQ. The remaining 14 species were used in all mixes.

Step	Temperature, °C	Duration, min	Cycles
Initial denaturation	95	3	1, hold
Denaturation	98	0.25	Ranged from 3 to 8
Annealing	59	0.5	depending on
Extension	72	1.5	sample
Final extension	72	5	1, hold

Table 2.1: PCR protocol for the WGS library preparation.

2.2.3 WGS sequencing library creation

DNA shearing was performed using the Covaris S2 sonicator (Covaris®) with the following settings: duty factor = 10%, intensity = 2.5, cycles/burst = 200, temperature = 6°C, total time, sec = 45. Size selection was performed on the sheared products with Ampure XP beads (Agencourt) to maintain insert size around 450 base pairs. Illumina sequencing libraries were prepared by ligating custom Illumina Truseq adapters with dual barcoding (10 base pairs) using the KAPA Hyper Prep Library Preparation kit (KAPA Biosystems, Inc.). To increase library yield, additional library amplification was performed with KAPA HIFI HotStart ReadyMix using the PCR protocol described in Table 2.1. To enable balanced pooling, sequencing libraries were quantified in duplicate by real time PCR using the KAPA SYBR®FAST qPCR kit. Quantification reactions were performed on a LightCycler®480 (Roche) using a dilution series of PhiX control library (Illumina) as standard [210]. After pooling the libraries, the final pool was quantified again using the same method to enable optimal loading of the flow cell.

2.2.4 16S sequencing library creation

Previously published [211] Primers and PCR-protocol for the amplification of V3-V4 region of the 16S rRNA were used. Illumina sequencing libraries were prepared by ligating custom Illumina Truseq adapters with dual barcoding (10 base pairs) using the KAPA Hyper Prep Library Preparation kit (KAPA Biosystems, Inc.).

2.2.5 DNA sequencing

Sequencing of WGS and 16S libraries was performed on the MiSeq®sequencer (Illumina) using v3 sequencing reagents according to the manufacturers' protocol with approximately 5% of PhiX control. This yielded one paired-end dataset with a read length of 299 bp per sample.

2.2.6 Bacterial genomes assembly

Sequencing reads for each bacterium were preprocessed using the Flexiprep quality control pipeline[1]. Post-QC reads were assembled by SPAdes Genome Assembler [212] with default settings.

2.2.7 Regression analysis

k-mer counting was performed using command count of the kPAL toolkit [213] with *k* set to 11. In case of the absence of the alternative DNA stand, *k*-mer profiles were balanced with balance command of the kPAL toolkit. Linear regression was done using the scikit-learn package for Python [214] with the *fit_intercept* parameter set to "False". The model training and prediction was performed using 5-fold Cross Validation.

2.2.8 Analysis using Centrifuge

Centrifuge is a popular tool that allows for fast classification of reads in a metagenomic sample using comparison of *k*-mers derived from each read to an indexed database. Centrifuge performs classification for all reads in a metagenomic sample independently using the following algorithm. For each read it creates a classification tree by pruning the taxonomy and only retaining taxa (including ancestors) associated with *k*-mers found in that read. Each node is weighted by the number of *k*-mers mapped to the node, and the path from root to leaf with the highest sum of weights is used to classify the read. A fast and effective comparison is achieved using the genome indexing technique, which is based on the Burrows-Wheeler transform [112] and the Ferragina-Manzini index [113]. To perform taxonomy assignment, Centrifuge requires an indexed database which is based on the reference database and its associated phylogenetic tree. A number of popular and regularly updated premade indexed databases are available on the Centrifuge website[2]. It is also possible to create a custom Centrifuge indexed database.
Metagenomic mixes samples were subjected to a QC-check using FastQC[3] (version 0.11.7). Leftover adapter removal and quality trimming of the reads was performed with cutadapt [95] (version 1.16, using options *--trim-n*, *--minimum-length* = 50 and *--quality-cutoff* = 20). The number of reads before and after each aforementioned step can be found in supplementary Table S1. High quality pairs of overlapping reads were merged with FLASH [215] (version 1.2.11, using option *--max-overlap*=300). For the subsequent taxonomic classification with Centrifuge, both merged reads and pairs of non-merged reads were used.

[1] Available online at http://biopet-docs.readthedocs.io/en/latest/pipelines/flexiprep/
[2] ftp://ftp.ccb.jhu.edu/pub/infphilo/centrifuge/data
[3] Available online at:http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Post-QC reads were analysed with Centrifuge (version 1-0-2-beta, default settings). Three different reference databases were used for the analysis: RefSeq database of complete genomes of bacteria and archaea [216] (downloaded as premade in April 2018 Centrifuge index); GreenGenes 16S sequences database (downloaded in June 2018) and SILVA 16S sequences database (downloaded June 2018). In order to make the content of reference databases comparable, sequences marked as eukaryotic were removed from SILVA database. Results obtained by Centrifuge were analysed using the Pavian interactive browser application [217].

2.2.9 Analysis using MG-RAST

MG-RAST is a web-based tool that allows the user to upload sequences and their metadata and download the analysis results. The MG-RAST pipeline creates a metagenomic profile by extracting rRNA and protein coding sequences. Gene calling is performed by the FragGenescan [119] algorithm, predicted protein sequences are clustered using UCLUST [101]. Potential rRNA genes are identified using BLAT [107] against a reduced version of the SILVA database and clustered with UCLUST. From each obtained cluster one representative sequence (the longest one) is chosen for the comparison with a reference database (M5nr58 [218] for proteins and combination of SILVA59, GreenGenes42 and RDP41 for rRNA analysis) using BLAT. All sequences from a particular cluster are assigned to the same taxonomic group as the clusters' representative. Thus, only rRNA genes and functional genes are used for the analysis of the metagenome, and the reads assignments are not independent. This strategy allows MG-RAST to perform taxonomic and functional profiling of metagenomic data. Finally, MG-RAST supports different metagenomic datatypes: genomic (including WGS and 16S) and transcriptomic. It also considers the metagenome origin, sequencing platform and many other features to tune the pipeline for a specific task. Raw reads of bacterial mixes samples were submitted to the MG-RAST Metagenomics analysis server under project number 85582. Paired reads merging and quality control was performed as part of the standard MG-RAST pipeline.

2.2.10 Taxa abundance estimation and results evaluation

Since the 16S amplification product has the same length among all bacterial taxa, no correction for genome length is needed when estimating relative abundances of the taxa. For the WGS samples however, normalization of read counts is required because of the differences in genome lengths. In order to perform correct taxa abundance estimations for taxonomic ranks higher than species, it is important to know how many reads assigned to that taxon belong to each species within the taxon. Both tools, Centrifuge and MG-RAST, assign reads to a node in the phylogenetic tree. Thus, reads assigned to a particular genus, for example, might belong to each of the species included to that genus as well as to the genus itself,

without species annotation. The main assumption of our approach for the estimation of taxa abundances is the following: all reads, assigned to the node higher than species level (regardless of whether or not they have species annotation), will be distributed among the species belonging to that node the same way as the reads with known species annotation. If the estimated abundances for species were known (in case of taxonomic annotation with Centrifuge), the procedure is trivial. When performing the analysis with MG-RAST the reads are classified only up to the genus level. In that case an equal distribution of reads among the species belonging to the particular genus was assumed.

2.2.11 Statistical and correlation analysis

Correlation analysis was performed using the Pearson correlation coefficient, pair wise comparisons were performed using the two-sided Mann-Whitney U test [219] and False Discovery Rate (FDR, a statistical approach used in multiple hypothesisto correct for multiple comparisons) control was performed using the Benjamini-Hochberg procedure [220]. We used the ratio of properly predicted taxa to all taxa predicted at that rank as a measure for the precision. Sensitivity was calculated as the ratio of properly predicted taxa to all taxa that were supposed to be present in the sample at that rank. F-scores (a measure of accuracy that considers both precision and sensitivity) were calculated as described in [221].

2.3 Results and Discussions

2.3.1 Individual bacterial genomes assembly

We sequenced and assembled the genomes of all 15 selected skin-associated bacteria individually. The total length of the assembly for each species was comparable to the length of the species references (Table 2.2 and section S1 of the Supplementary materials). For one species (*A. lwoffii*) there was no reference sequence available. Obtained assembly lengths as well as the DNA concentration measured for each bacterium were used to create four metagenomic mixes: one with equal and three with exponential ($\lambda = 1/6$, $\lambda = 1/2$ and $\lambda = 5/6$) distribution of bacterial species abundances. Taxa abundances were ordered from high to low as shown in Fig. 2.1.

2.3.2 Estimation of reference abundances

In order to estimate an abundance of an organism in terms of genome copies, the length of the genome and the lengths and (relative) copy numbers of any plasmids needs to be known. In the absence of a strain-specific reference sequence, *de novo* assembly of a single organism can be used to obtain these data [222]. In most common approaches [223], the coverage (and thereby the copy number) of contigs (see Supplementary Fig. S1) is not considered when estimating an assembly length, which leads to an inaccurate estimation of the organisms' genome length and thus influence the accuracy when creating bacterial mixes (see Supplementary Fig. S2 for a step-by-step explanation). Other factors, such as inaccuracy in DNA concentration measurement or mixing, can also lead to different abundances in the final bacterial mixes from those intended.

Since the content of all our metagenomic mixes is known and individual assemblies of all bacterial species were available, the intended distribution of bacterial abundances in the metagenomic mixes could be verified using the following approach. We used *k*-mer counts as a proxy for the number of genomes present in a pure (unmixed) sample. Using these counts, we are able to infer the relative contributions to a mixture. We use randomly chosen *k*-mers from the pure samples as profiles for the organisms, the same *k*-mers are used to make a profile of the mix and by linear regression, we estimate the contribution of each profile and thereby the contribution of each organism to the mix. For a more detailed description and a motivational example, see Section S1 and Figure S2 of the Supplementary materials. We calculated the 11-mer profiles for each bacteria using the contigs obtained after individual genome sequencing and assembly. Since profiles were calculated using contigs, we compensated for the absence of the reverse-complement DNA strand. We also calculated the 11-mer profiles of the WGS datasets of each of the metagenomic mixes, in these cases strand balancing was not applied. The 11-mer profiles were used to build a linear regression model in which the individual bacterial *k*-mer counts were

treated as independent variables and the k-mer counts of the metagenomic mix served as dependent variable.

Bacteria	Number of contigs	Accession number	Reference length, Mb	Assembly length, Mb
Acinetobacter johnsonii ATCC 17969	206	NZ_CP010350.1	3.51	3.88
Acinetobacter lwoffii ATCC 15309	180	NA	NA	3.44
Corynebacterium jeikeium ATCC 43734	234	NC_007164.1	2.46	2.6
Corynebacterium urealyticum ATCC 43042	99	NC_010545.1	2.37	2.35
Moraxella osloensis NCTC 10145	89	CP014234.1	2.43	2.58
Propionibacterium acnes ATCC 6919	26	NC_017550.1	2.49	2.55
Pseudomonas aeruginosa ATCC 10145	99	NC_002516.2	6.26	6.35
Staphylococcus aureus ATCC 29213	45	NZ_CP009361.1	2.78	2.72
Staphylococcus capitis ATCC 27840	52	NZ_CP007601.1	2.44	2.6
Staphylococcus epidermidis ATCC 12228	142	NC_00446	2.5	3.3
Staphylococcus haemolyticus ATCC 29970	770	NC_007168.1	2.69	2.86
Staphylococcus saprophyticus ATCC 15305	351	NC_007350.1	2.15	1.89
Streptococcus piogenes ATCC 19615	65	NZ_CP008926.1	1.84	1.82
Staphylococcus xylosus ATCC 29971	97	NZ_CP008724.1	2.52	2.74
Streptococcus mitis LMG 14552	49	NC_013853.1	2.76	2.83

Table 2.2: Bacterial species used for metagenomics mixes.

To verify the intended distribution of bacterial abundances in the metagenomic mixes, we use *k*-mer counts as a proxy for the number of genomes present in a pure (unmixed) sample. Using these counts, we are able to deconvolute a mixture. We use randomly chosen *k*-mers from the pure samples as profiles for the organisms,

the same *k*-mers are used to make a profile of the mix and by linear regression, we estimate the contribution of each profile and thereby the contribution of each organism to the mix. For a more detailed description and a motivational example, see section S1 and Figure S2 of the Supplementary materials. We calculated the 11-mer profiles for each bacterum using the contigs obtained after individual genome sequencing and assembly. Since profiles were calculated using contigs, we compensated for the absence of the reverse-complement DNA strand. We also calculated the 11-mer profiles of the WGS datasets of each of the metagenomic mixes, in these cases strand balancing was not applied. The 11-mer profiles were used to build a linear regression model in which the individual bacterial *k*-mer counts were treated as independent variables and the *k*-mer counts of the metagenomic mix served as dependent variable.

Since *k*-mer counts within one profile might be correlated, which violates the condition for using the regression analysis, we did not analyse the complete profile of 4,194,304 possible 11-mers. Instead we performed 1,000 iterations, in each iteration choosing 10,000 random *k*-mers and performing the regression analysis on that subset of *k*-mers. Thus, for each organism we got 1,000 estimations of its abundance in each mix. The result of this analysis is presented in Figure 2.1. Each boxplot shows the distribution of the organisms' abundances obtained from the regression analysis. The median model fit of the cross-validated models (measured using the R^2 coefficient of determination) for each mix was larger than 0.95, accuracy of the prediction (also measures using the R^2 but on the data that did not participate in the model training) ranged from 0.80 to 0.92 depending on the mix.

The regression analysis confirmed the distribution of bacterial abundances we aimed for (uniform distribution turning into the exponential one), though for some species (e.g., *S. haemoliticus* and *P. aeruginosa*), slight positive or negative deviations from the anticipated values were found. This can be caused by a number of factors such as inaccuracy in the DNA concentration measurement or DNA mixing, presence of large amounts of non-chromosomal DNA (e.g., plasmids) in the pool of bacterial DNA or inaccuracy in bacterial genome size estimation.

We use the results of this analysis as reference abundances for the experiments done in section 2.3.5.

2.3.3 Analysis of bacterial mixes using Centrifuge and MG-RAST

The mixes were sequenced on the Illumina MiSeq using WGS (samples EQ_WGS, EXP16_WGS, EXP12_WGS and EXP56_WGS) and 16S for V3-V4 region (samples EQ_16S, EXP16_16S, EXP12_16S and EXP56_16S) protocols. Information about read counts and QC statistics for each obtained dataset can be found in Supplementary table S1.

WGS and 16S samples obtained from our four metagenomic mixes were analysed with Centrifuge using the RefSeq complete bacterial genomes database. We per-

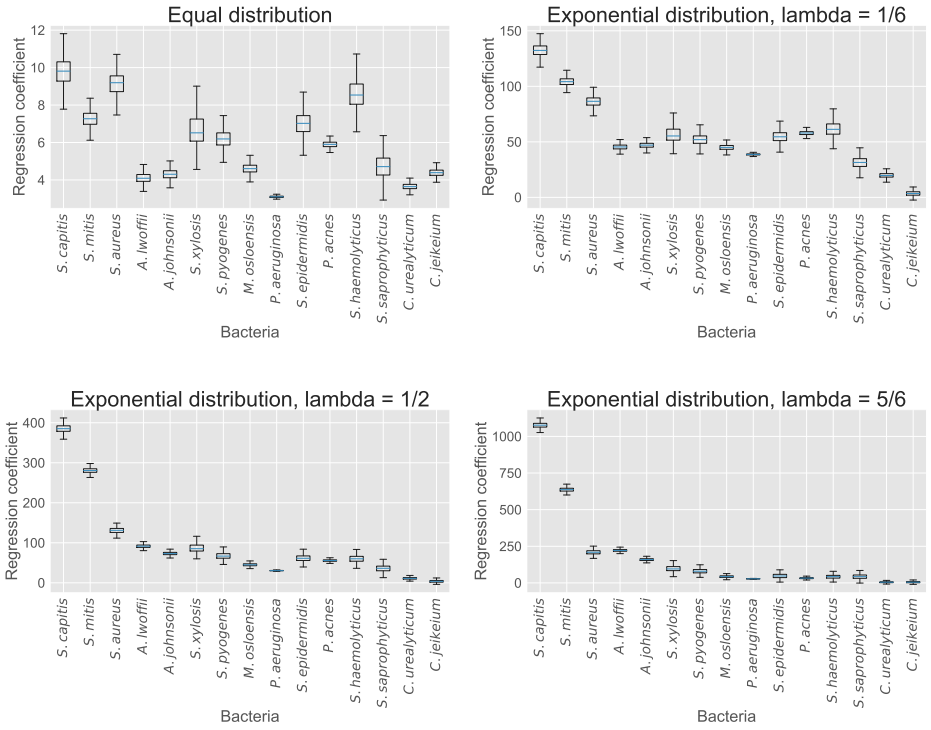

Figure 2.1: Regression analysis performed for metagenomic mixes to estimate relative abundances. Results for each mix are shown on a separate plot. Each boxplot represents the distribution of regression coefficients (vertical axes) obtained for the particular organism (horizontal axes), thus representing the distribution of bacterial abundances within that particular mix.

formed additional analysis for 16S samples using Centrifuge with the GreenGenes and SILVA reference databases.

All eight datasets (four WGS and four 16S) were submitted to the MG-RAST Metagenomics analysis server under project number 85582. RefSeq and GreenGenes databases provide taxonomic annotation down to the species level, while SILVA database as well as the databases used by MG-RAST are restricted to the genus level. Since the NCBI taxonomy and the taxonomy used by MG-RAST were different at the order level for our set of bacteria, we excluded annotation at the order level from further analysis.

2.3.4 Profiling accuracy without considering relative abundances

Because the content of the metagenomic mixes is known, we can verify how many of the reported taxa on each taxonomic rank are correct (true positive counts), how many are incorrect (false positive counts) and how many are missed (false negative counts).
Using these counts, both precision and sensitivity can be calculated. A perfect prediction is made if both precision and sensitivity equal one. As can be seen in Figure 2.2, both precision and sensitivity tend to increase in all cases with increasing taxonomic rank. For all 16S datasets analysed with Centrifuge, we observe that precision never reaches its maximum value, while for WGS datasets analysed with Centrifuge precision reaches its maximum already at the genus level. Interestingly, for 16S datasets analysed with MG-RAST, precision reaches its maximum at the genus level, but the sensitivity does not increase any further. For WGS datasets analysed with MG-RAST, sensitivity reaches its maximum already at the family level.
The accuracy of the classifications can be expressed using the F-score, which is calculated using precision and sensitivity. We tested whether the F-scores differed significantly for each pair-wise comparison using the Mann-Whitney U test and the Benjamini-Hochberg procedure for FDR control. The full table of p-values can be found in Supplementary Table S2, a summary of the results is shown in Figure 2.3. In most cases, the F-scores differ significantly when comparing WGS to 16S. At the same time, when comparing WGS datasets with different tools, a significant difference was observed only at the genus level.

2.3.5 Abundance assignment accuracy

Both Centrifuge and MG-RAST provide read counts for each reported taxon. We considered only reads that were assigned to the expected taxa and compared their relative abundances to the reference abundances.
Only Centrifuge, when using either the RefSeq or GreenGenes database, reported the taxonomic assignment down to the species level. In Figure 2.4, each metagenomic mix is shown as a separate graph with species listed on the horizontal axes and their relative abundances shown on the vertical axes. The black line represents the intended distribution of species abundances. The dark green line shows the mean reference abundances with the light green area representing ± 3 standard deviation around those means. The blue and red lines show the relative abundances obtained for 16S and WGS datasets respectively, with the solid blue line for the 16S analysis done using the RefSeq database and the dashed blue line using the GreenGenes database. As can be seen in Figure 2.4, the analysis of 16S data results in a considerable overestimation of abundance of *A. johnsonii*. Centrifuge failed to identify *A. lwoffii*, since there is no complete genome of that bacterium in the RefSeq database and it did not report any significant presence of *C. jeikeium* in

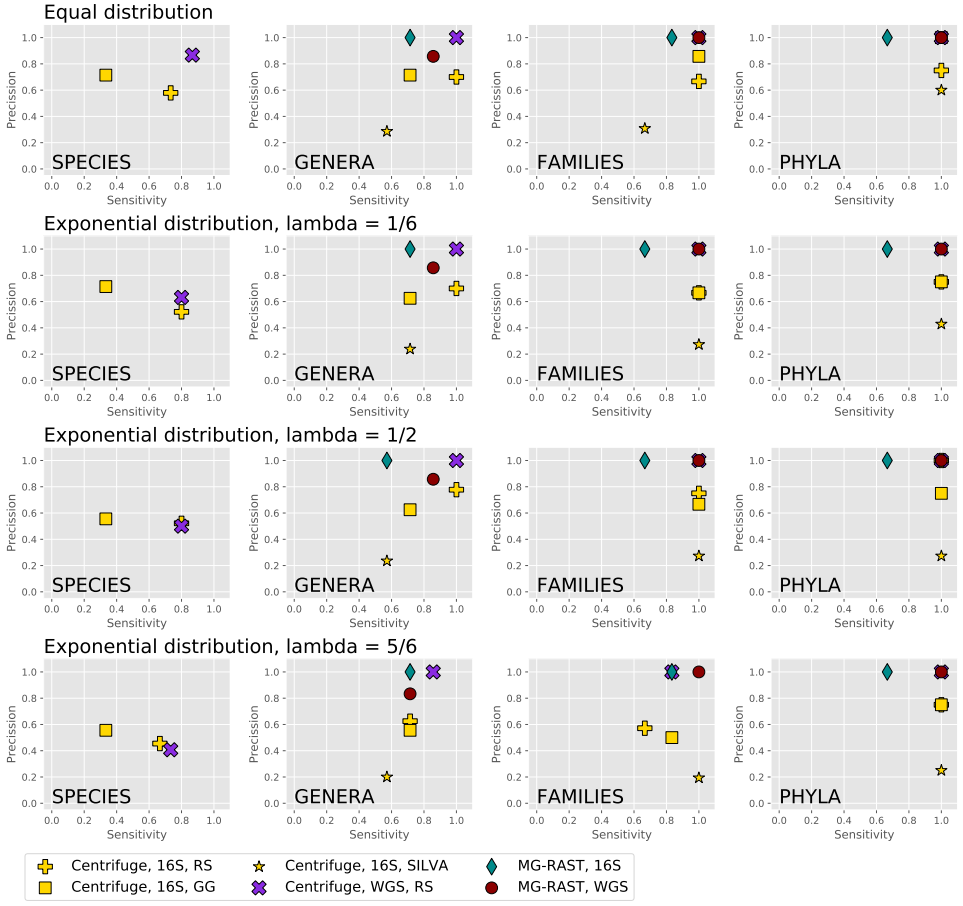

Figure 2.2: Precision vs. sensitivity of different profiling approaches. Results for each mix and taxonomic rank are shown separately, with sensitivity on the horizontal axes and the precision on the vertical axes. Each shape represents a combination of method, data type and reference database. RS - RefSeq database, GG - Greengenes database, S - SILVA database.

the exponentially distributed metagenomic mixes. Analysis of the 16S datasets using the GreenGenes database reported overestimated values for *S. epidermidis* and *A. johnsonii* and did not report the presence of nine out of fifteen bacteria because of their absence in the GreenGenes database.

We repeated the same analysis on three higher taxonomic ranks: genera, families and phyla. For all these three taxonomic levels we analysed the results of Centrifuge (Figure 2.5) and MG-RAST (Figure 2.6). As can be seen in Figure 2.5, the Centrifuge analysis of 16S datasets using different reference databases provided a similar biased

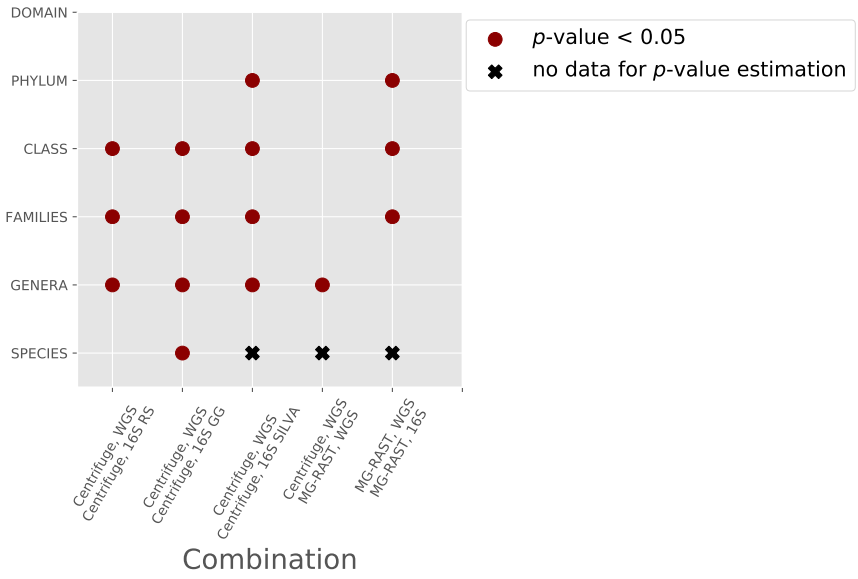

Figure 2.3: Comparison of F-scores (combination of precision and sensitivity) obtained from all four mixes for different combinations of methods, data type and databases. Red dots indicate a *p*-value below 0.05. Combinations of methods, data type and database are shown on the horizontal axis, taxonomic levels are shown on the vertical axis. RS - RefSeq database, GG - Greengenes database, S - SILVA database. Please note that data points are connected only to visualize the various types of distributions.

output, mostly due to an overestimation of the abundance of the Acinetobacter genus, Moraxelaceae family and Proteobacteria phylum. The dissimilarity with the reference abundances is especially pronounced at the phylum level. Results obtained for the WGS datasets with Centrifuge were concordant with the reference abundances with slight deviation for Acinetobacter genus, Moraxelaceae family and Proteobacteria phylum (Figure 2.5). It is interesting to note, that these taxa were also the major reason for disagreement between results obtained by Centrifuge for 16S datasets and reference abundances.

The results obtained for different 16S datasets by MG-RAST were not consistent (as is the case for Centrifuge) up to the phylum level. As can be seen in Figure 2.6, analysis of 16S datasets with MG-RAST reported many disagreements with reference abundances. The reasons of those disagreements are dataset- and taxonomy rank-specific. Results reported by MG-RAST became more or less consistent only at the phylum level, where they followed the same trend: overestimating the abundance of Firmicutes relative to that of Proteobacteria.

Abundances obtained after analysis with MG-RAST of WGS datasets were also following the reference results closely. There were, however, slight deviations from

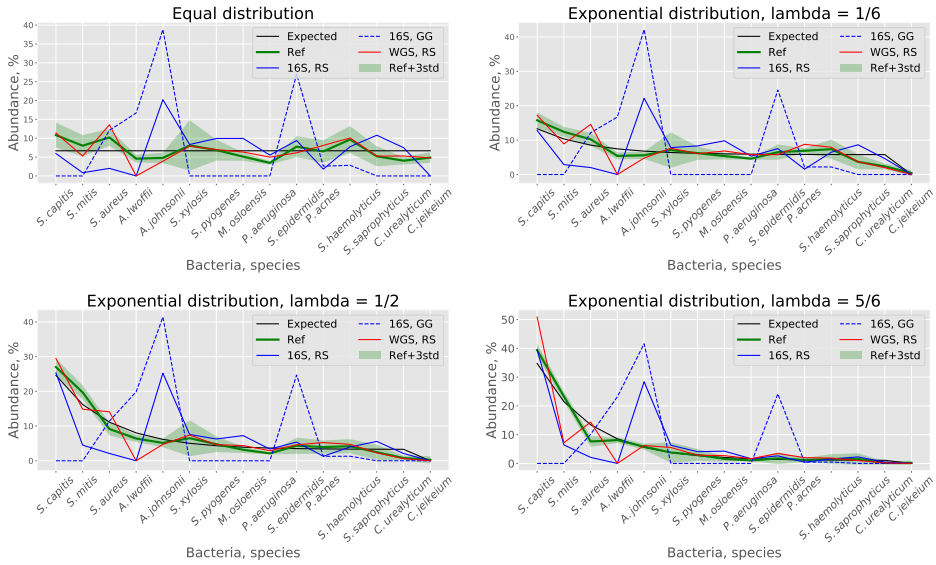

Figure 2.4: Comparison of relative abundances reported by Centrifuge (using two different reference databases) for WGS and 16S with relative abundances obtained from the regression analysis. Results for each mix are shown separately with the species' names on the horizontal axes and the relative abundance on the vertical axes. Ref - reference abundances, RS - RefSeq database, GG - GreenGenes database. Please note that data points are connected only to visualise the various types of distributions.

the reference abundances. These deviations were, like the results for 16S datasets, specific to taxonomy-rank and dataset.

In order to quantify the dissimilarity among the abundances provided by the different methods, datasets, reference databases and the results of regression analysis we calculated the absolute differences in abundances for each particular dataset and taxonomic rank. The averages of these values (from here on called the error rate) are reported in Figure 2.7. For the analyses of 16S datasets it is interesting to note that for Centrifuge the average error rate grew with the increase of the taxonomic rank in general. This was not the case for the error rate obtained for the 16S datasets using MG-RAST. We tested whether the average errors differed significantly for each pair-wise comparison using the Mann-Whitney U test and the Benjamini-Hochberg procedure for FDR control. The full table of p-values can be found in Supplementary Table S3, a summary of the results is shown in Figure 2.8. This analysis demonstrates that for all taxonomic levels the error rates in the abundance estimations provided by the analysis of 16S datasets (regardless of the method or reference database) are significantly different (higher) compared to the abundances reported for WGS datasets. We did not observe any significant difference in average error rate between

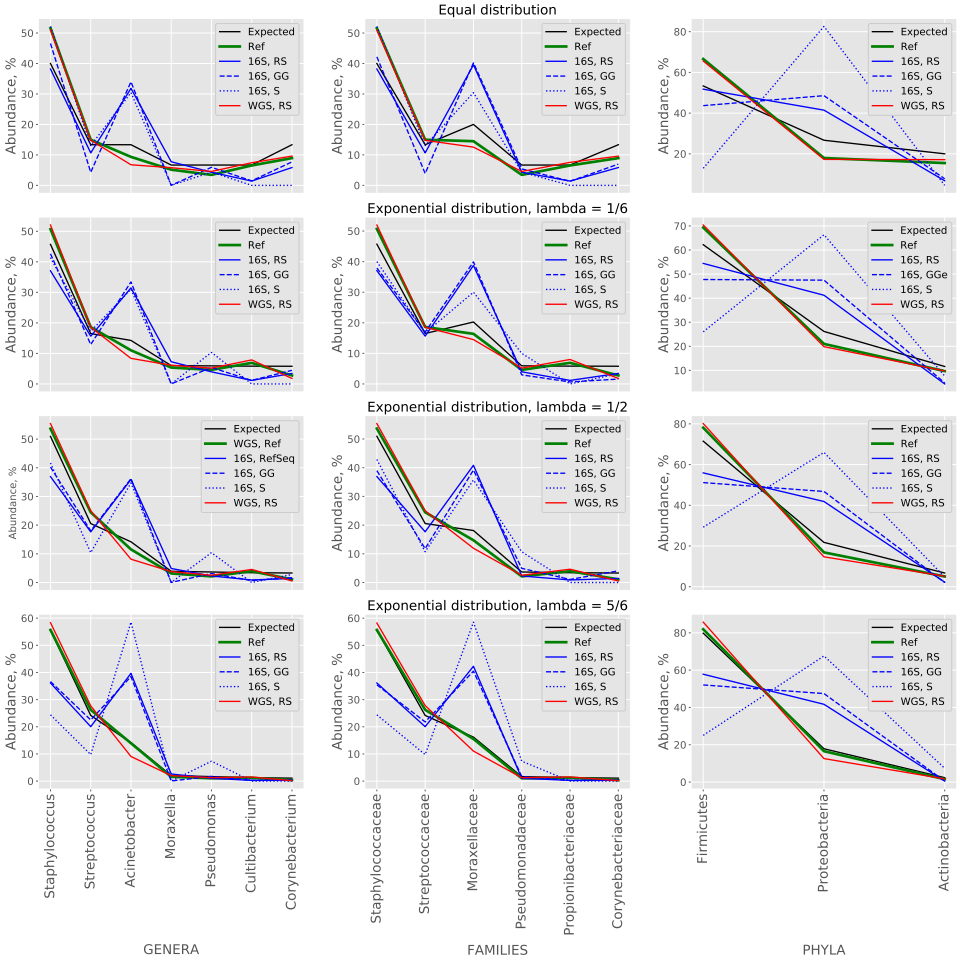

Figure 2.5: Comparison of relative abundances reported by Centrifuge (using three different reference databases) for WGS and 16S datasets on genera, orders and phyla levels with relative reference abundances. In the above grid of figures each row indicates the mix and each column indicates the taxonomic level. In each figure, the taxa are shown on the horizontal axes and the relative abundances are shown on the vertical axes. Ref - reference abundances, RS - RefSeq database, GG - GreenGenes database, S - SILVA database.

WGS datasets analysed with Centrifule and MG-RAST.

We compared the error rates reported by Centrifuge when using the three different 16S reference databases. Error rates observed in the analysis with RefSeq and GreenGenes databases were similar. Running the Centrifuge analysis using the SILVA database reported a much higher error rate. That might be a direct consequence of

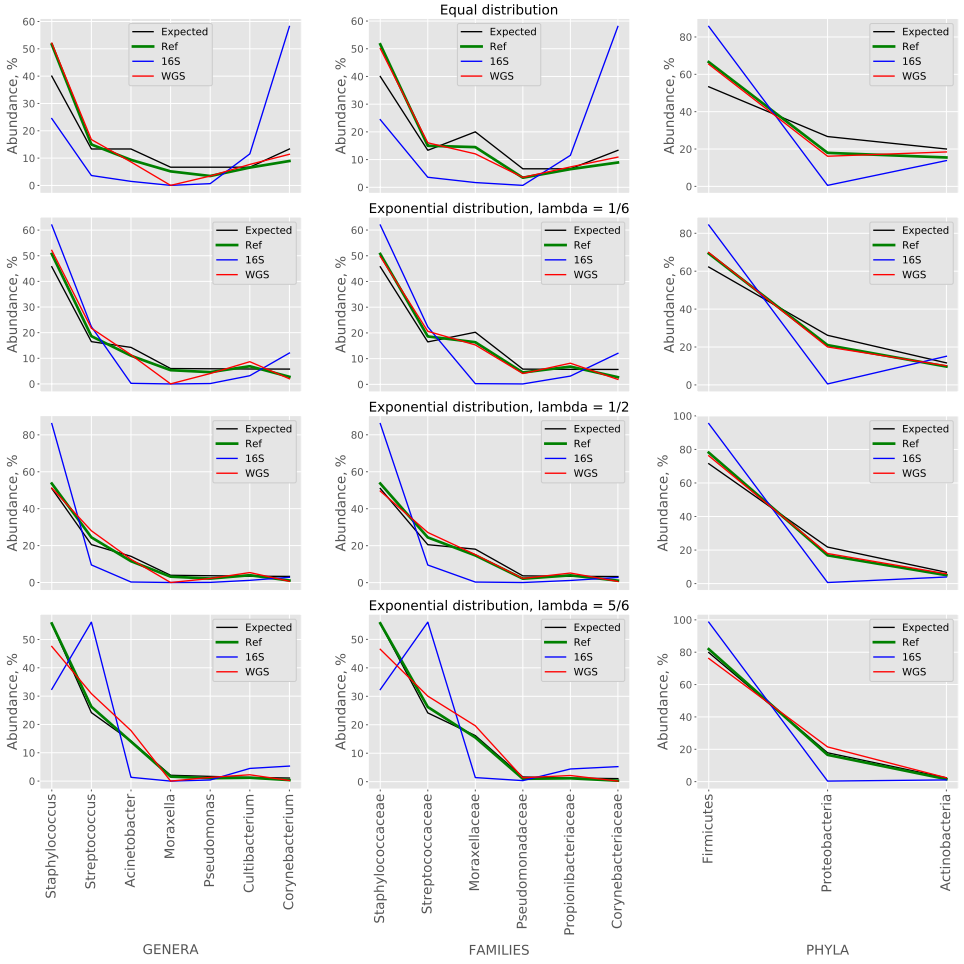

Figure 2.6: Comparison of relative abundances on reported by MG-RAST for WGS and 16S datasets on genera, orders and phyla levels with relative reference abundances. In the above grid of figures each row indicates the mix and each column indicates the taxonomic level. In each figure, the taxa are shown on the horizontal axes and the relative abundances are shown on the vertical axes. Ref - reference abundances

taxonomic annotation done using the SILVA database where a smaller proportion of reads was assigned to the expected taxa in comparison to other reference databases (see the section 2.3.4).

We also evaluated the similarity among the abundances obtained by employing distinct methods and databases using a correlation analysis. In Figure 2.9 the results of these comparisons are presented as a series of heatmaps.

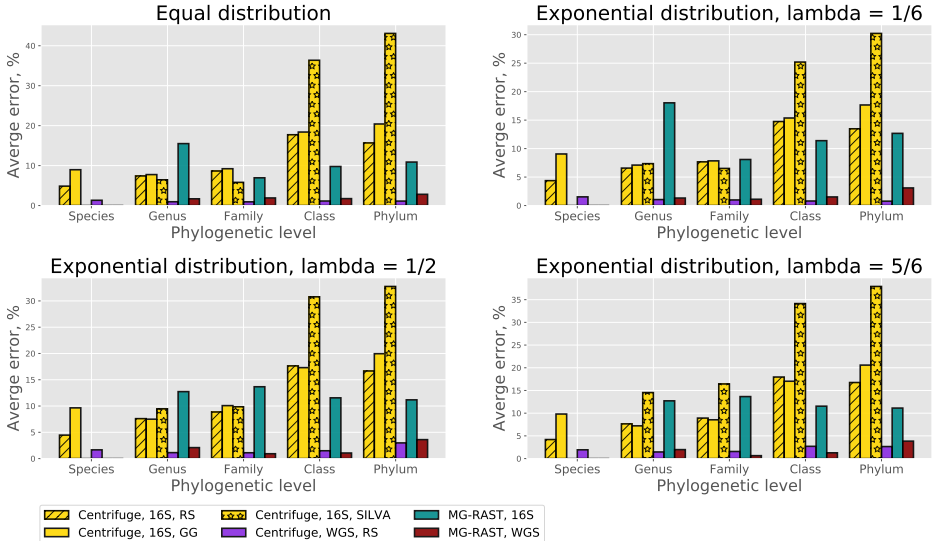

Figure 2.7: Average error between the abundances reported by the regression analysis and results obtained for different datasets, reference databases and tools. Results for each metagenomics mix are shown separately. Bar colours represent the tool, data type and reference database combination used for the analysis. The error rate (shown on the vertical axes) was calculated for each taxonomic rank (shown on the horizontal axes) separately. RS - RefSeq database, GG - GreenGenes database, SILVA - SILVA database

As can be seen from Figure 2.9, abundances obtained by the analysis of WGS data (Centrifuge and MG-RAST) for all datasets at all taxonomic levels positively correlate with reference abundances. Correlation of 16S analysis obtained using Centrifuge with the reference abundances becomes worse at higher taxonomic levels, which is the opposite for the 16S data results obtained using MG-RAST. The 16S data analyses obtained for Centrifuge and MG-RAST do not demonstrate positive correlation with each other.

2.4 Conclusions

In this study we created a series of bacterial mixes with known content in order to investigate which type of metagenomics data and reads assignment strategy yields better taxonomic classification. For each mix we generated WGS and 16S sequencing datasets and analysed them using Centrifuge with RefSeq, GreenGenes and SILVA reference databases and the MG-RAST metagenomics analysis server with M5nr and M5nra reference databases. We compared the results of all analysis done with Centrifuge and MG-RAST to the reference abundance profiles obtained from a regression *k*-mer-based regression analysis.

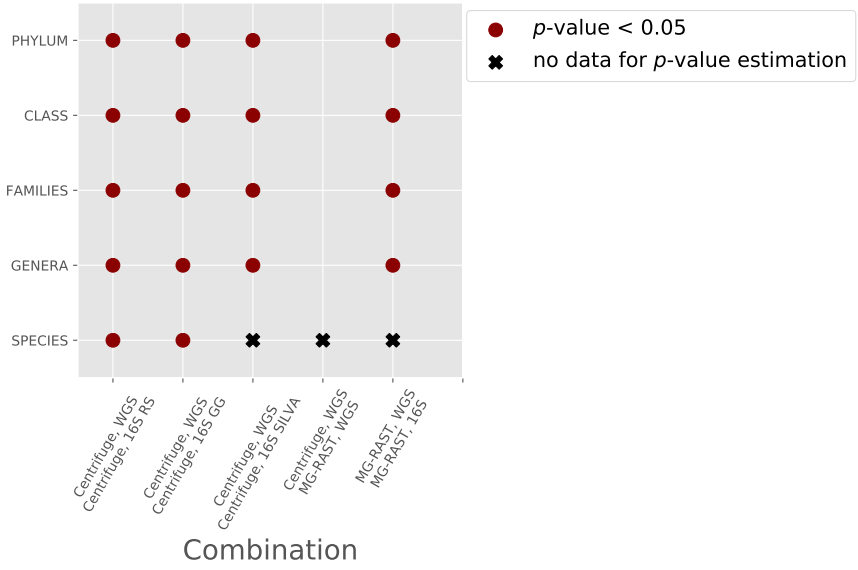

Figure 2.8: Comparison of average errors obtained from different mixes for different combinations of methods, data type and databases. Combinations of methods, data type and database are shown on the horizontal axis, taxonomic levels are shown on the vertical axis. Red dots indicate a *p*-value below 0.05. RS - RefSeq database, GG - GreenGenes database, SILVA - SILVA database

The results from both Centrifuge and MG-RAST show that WGS datasets provide much more accurate results in comparison to 16S-based methods. The analysis of WGS data displayed better coverage of all taxa expected to be present in the mixes on all phylogenetic levels, reaching the maximum accuracy already at the genus level for Centrifuge and at the family level for MG-RAST. On the other hand, results obtained for 16S-based data were often missing several taxa and/or had very high false-positive rate. Centrifuge analyses based on the 16S datasets were suffering from low precision, while MG-RAST analysis of the 16S datasets had low sensitivity. Abundance profiles obtained from WGS demonstrated much less disagreement with the expected abundances in comparison to the abundance profiles based on 16S data. This was shown using two different measurements: the average (per taxonomy rank) absolute difference between abundance profiles and by a correlation analysis. For 16S datasets analysed with Centrifuge, the deviation from the reference abundances, introduced at the species/genus levels, propagated further up the taxonomy which led to a greater difference with the expected outcome on the higher taxonomic ranks as well. In contrast, the analysis of 16S datasets performed by the MG-RAST pipeline demonstrated greater differences with the reference abundances on the lower taxonomic ranks in comparison with the higher ones. Our correlation anal-

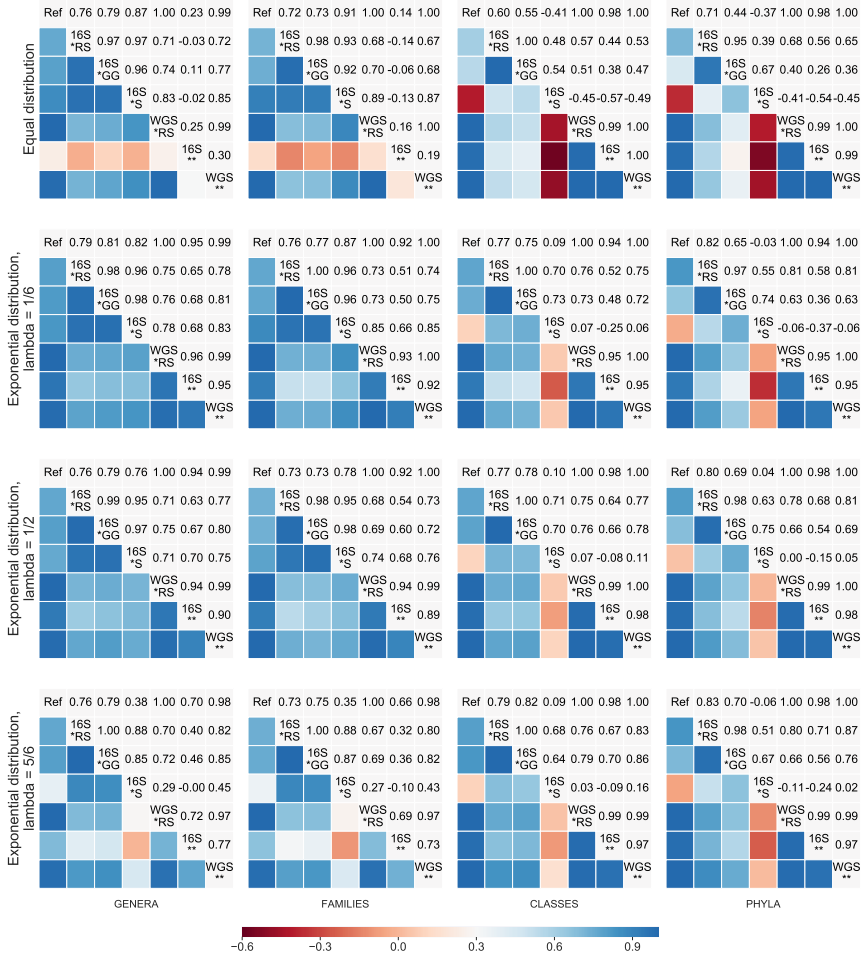

Figure 2.9: Correlation between the abundance profiles obtained for different combinations of method, datasets and reference database on distinct taxonomic levels. In the above grid of figures each row indicates the mix and each column indicates the taxonomic level. The combination of analysis type, dataset and reference database are shown on the main diagonal of the heatmap, with the lower triangle representing the correlation shown in colors and the upper triangle demonstrating the same data in the numeric representation. Ref - reference abundances, * - Centrifuge, ** - MG-RAST, RS - RefSeq database, GG 435 - GreenGenes database, S - SILVA database.

ysis shows that the agreement between the MG-RAST results of 16S datasets and reference abundances was growing with increasing taxonomic level.
Both tailor-made 16S databases (GreenGenes and SILVA) did not perform better than the RefSeq database when analysing 16S datasets using Centrifuge. The Centrifuge results using RefSeq and GreenGenes databases were correlated with a correlation

coefficient higher than 0.95 for all 16S datasets on each taxonomic rank starting with genus.

We conclude that WGS data is preferable for the study of metagenomic data, especially when the correct inhabitant abundances are required. We could not determine which of the explored methods for the taxonomic assignment of the WGS data provides a more accurate outcome. Centrifuge, however, has minor advantages in comparison to MG-RAST, such as a faster, deeper and slightly better reads classification, the possibility of local installation and use of custom databases and a more flexible tuning of the tools' settings. Among the investigated techniques for 16S metagenomic data analysis, MG-RAST demonstrated slightly better results in both reads assignment and abundance estimation, albeit only at higher taxonomic ranks. As previously quoted, "the capacity of WGS data of microbiomes to aid in forensic investigations by connecting objects and environments to individuals has been poorly investigated". In light of this, our results are especially important, as they demonstrate the inefficiency of routine 16S data to produce the accurate taxonomical profiling.

The synthetic metagenomes created in our study is restricted to DNA of bacteria that inhabit skin surface - a logical target for forensics analysis. However, human skin is also the environment with one of the most within- and between-individual diverse microbiota on the human body. The benchmark we created is rather small and simple as the diversity of microbial species living on the human skin surface is much larger than only 15 species [224]. The significant inaccuracy of the results obtained for 16S data in comparison with those for WGS data on a small and simple set of benchmarks can possibly question the accuracy of the previous 16S-based forensic studies, at least those done on skin-associated microbial communities.

2.5 Author Statements

2.5.1 Funding information

This research is financed by a grant number 727.011.002 of the Netherlands Organisation for Scientific Research (NWO).

2.5.2 Authors' contributions

Conceptualization LK and JFJL; Methodology RHdL, MtR; Resources MEMK; Software LK and NP; Investigation, Visualization, and Writing original draft LK; Supervision JFJL, PdK and HM; Funding Acquisition JFJL; Reviewing and editing LK, JFJL, HM, NP, PdK, RHdL and MEMK.

2.5.3 Acknowledgements

We would like to thank Guy Allard for the support with the assembly of bacterial genomes and Louk Rademaker for the feedback on this manuscript.

2.5.4 Conflicts of interest

The authors declare that there are no conflicts of interest.

2.6 Data Availability

- Reads obtained after the individual sequencing of each selected bacterial species and used for the genome assembly were upload to the Sequence Read Archive under the study SRP159200.
- Sequencing reads of the metagenomic mixes as well as the results of the analysis performed by MG-RAST can be downloaded from the MG-RAST server (project number 85582).
- The summary of results obtained using Centrifuge as well as the supplementary materials for this research are deposited on Figshare: https://doi.org/10.6084/m9.figshare.c.4217672