Comparative performance of reference-based metagenomic tools to identify species-level taxa among families of bacteria: benchmarking *Mycobacteriaceae* and *Neisseriaceae*

Running head: Performance species-level metagenomic tools

Luke B. Harrison[a,#], Juba O. Ahmed[a] and Frédéric Veyrier[a]

Author Affiliations:

[a]Bacterial Symbionts Evolution. Centre Armand-Frappier Santé Biotechnologie. Institut national de la recherche scientifique. 531, boul. des Prairies. Laval (Québec) H7V 1B7. Canada.

#Address correspondence to Luke B. Harrison, luke.harrison@inrs.ca, luke.harrison@mail.mcgill.ca. Corresponding

Abstract (242 words)

Hypotheses concerning the ecology and evolution of bacteria commonly relate to the presence and abundance of species in various settings and conditions. Shotgun metagenomics may address these hypotheses, which previously relied on PCR or culture. However, the problem of determining the presence/absence of a given species of interest is not trivial, particularly when closely related species are present in the reference database or metagenomic sample. Reference-based methods to detect species-level taxa mostly rely on thresholding of aligned reads or mapped k-mers or derivative metrics like genomic coverage, and create a tradeoff between recall/completeness and precision/purity. Three methods for species-level profiling (YACHT, metapresence and sylph) have recently been published. Here we test the performance of these methods to detect related species of interest using simulated metagenomic samples from genomes in the families *Mycobacteriaceae* and *Neisseriaceae*, which contain closely related genomes. Among methods tested, metapresence, when used with

an alignment quality filter, and sylph offer the best overall performance. Sylph maintains high precision but requires a depth of coverage greater than approximately 0.1x to reliably detect a genome's presence. Metapresence has a lower limit of detection of hundreds of reads but this is balanced against relatively lower precision. Both methods are robust to the presence of reads from genomes outside the groups of interest. We demonstrate the application of these methods in two real-world datasets: a mycobacterial community in a drinking water system and the community of *Neisseriaceae* present in the human oral cavity.

Importance (149 words)

Detecting which bacterial species of interest are present in a given sample is fundamental to studies of microbial ecology and evolution, and to applied microbiology (e.g. clinical diagnostics). Culture-dependent and independent (e.g. PCR) approaches are increasingly complemented by metagenomic approaches, but methods to accurately identify specific low-abundance species-level genomes in a shotgun metagenomic sample are still being refined. Here we comprehensively test three methods (YACHT, metapresence, and sylph) using two simulated datasets of bacterial families, *Mycobacteriaceae* and *Neisseriaceae* that contain closely related species. Our simulations exploit natural genomic diversity to create a challenging benchmark. We demonstrate that metapresence and sylph perform well, with the former being well-suited to low-biomass host-associated datasets, and the latter with environmental metagenomic samples. This study is the first extensive benchmark of these methods for this use case, and demonstrates these methods can accurately identify closely related species of interest using an arbitrary reference database.

Introduction

Determining the presence or absence of a given species in an ecological niche can be accomplished through culture-dependent and culture-independent methods, frequently based on PCR amplification of marker genes with amplicon sequencing, for example of 16S rRNA genes, or *hsp65* mycobacteria-specific PCR[1]. Shotgun metagenomic sequencing has the potential to characterize the composition of microbial communities and identify the presence of bacterial species, irrespective of cultivability, and to concomitantly offer insight into the aggregate functional content of the microbial genomes[2,3]. Conceptually, these approaches can broadly be divided into reference-free methods that seek to produce meta-assembled genomes (MAGs) from de novo assembly of sequencing reads[3,4], and reference-based methods that rely on a database of genomes (which may also be MAGs) against which sequencing reads are mapped[2]. Most recently, with new reference-based profiling algorithms, shotgun metagenomics offers the promise of accurate species-level Identification in complex metagenomic samples, at low abundance[5–8]. This opens the door to the broad application of metagenomic methods as tools for the targeted identification of specific species-level taxa, among an arbitrary taxonomic group of interest, in both existing and prospectively collected samples without employing resource-intensive upstream wet-lab workflows (e.g. specific PCR with amplicon sequencing or bait-capture).

Identifying specific species present is relevant to test hypotheses concerning the ecology and evolution of microorganisms, and to interrogate the etiological agents of infectious diseases. The family *Mycobacteriaceae* is clinically relevant group of mostly environmental bacteria, whose precise niches are informed by both culture-dependent and culture-independent[9–11] methods, for which many questions remain[12]. More precise delineation of the habitats of mycobacteria may inform targeted interventions to reduce exposures and disease among vulnerable populations[12,13]. A second example is the tropism of species of the family *Neisseriaceae*, whose closely related members inhibit narrow symbiotic niches amongst their

vertebrate hosts[14], and whose evolutionary history is driven in part by niche-specific host interactions[15]. Previously, to address these hypotheses, researchers investigating these groups have employed techniques including cultures using enriched, selective and/or differential media, targeted PCRs and amplicon sequencing[16], and some metagenomics approaches[10,14].

Interrogating these questions using metagenomic sequencing, where testing hypotheses requires the discrimination of the presence and abundance of defined organisms at the species level, at a potentially very low abundance, is best approached using a reference-based method[6]. This metagenomic problem differs from the more general problem of taxonomic classification of sequencing reads or community profiling (e.g. microbiome composition), and we focus on the performance of these methods to determine the presence or absence of a set of (mostly) species-level genomes in a given sample, irrespective of overall community composition. However, this is not trivial[6] because bacteria share significant genomic sequence similarity through vertical evolution (e.g. phylogenetic relatedness) or horizontal evolution (e.g. gene transfer)[5,6,17]. For these reasons, and because bacterial species may be named and defined for historical or phenotypic reasons, despite having high average nucleotide identity (ANI), individual sequencing reads may map well to the genomes of multiple species[5,6,18]. Further, within a species, those organisms with open pangenomes may harbor significant variation in gene content relative to a given reference genome[19]. This complicates both the assignment of reads to a given taxon and the determination of which species are truly present/absent in a given metagenome[18]. Methods based on k-mer matching (e.g. Kraken2[20] and related algorithms) are known to yield false positive read read classifications at the species level, necessitating strict thresholding, and whose precision at the species level is generally low[5]. These problems may be further compounded in host-associated metagenomic studies due to low biomass, and/or incomplete host sequence removal[21].

These problems have been recognized, and several algorithms to improve the recall (ability to detect a genome that is truly present, also called sensitivity or completeness) and

precision (the fraction of genomes identified by the method as present that are truly present and not false positives, also called purity) of metagenomic tools have been published. These incorporate information such as spatial mapped read distributions, and read sequence divergence (e.g. KMCP[7]) or employ marker gene approaches (e.g. MetaPhlAn4[22]). More recently, several new algorithms for arbitrary database reference-based, species-level metagenomic profiling have recently been published. Among them, two are k-mer-based tools that use containment ANI (YACHT[8] and sylph[5]) and one is read alignment distribution-based (metapresence[6]). Here, we test the performance of these tools for the specific problem of determining which related species within a given taxonomic group are present within a metagenomic sample, regardless of the overall community composition. Rather than opting for existing simulated communities and tools (e.g. CAMI II challenges[23]), we explored performance of these methods on simulated microbial communities of closely related species in a single family, choosing taxa pragmatically (here *Mycobacteriaceae* and *Neisseriaceae*), that may contain species and species complexes of variable genetic interrelatedness. In an attempt to create a realistic, but challenging real-world benchmark, these datasets are expected to contain species with open pangenomes and significant HGT (in the case of naturally competent *Neisseriaceae*, HGT is particularly widespread[24]) where pairwise ANI is variable and may exceed 95%. We further enriched our simulation with natural genome evolutionary dynamics by exploiting the diversity of publicly available genome sequences to create our simulated datasets rather than relying only on reference genome sequences themselves. We do not compare these methods with marker-based approaches (e.g. Metaphlan4[22]), which rely on predefined taxonomic databases, instead limiting to methods permitting arbitrary reference databases.

These methods implicitly or explicitly use thresholds to define a given genome's presence: this creates a trade off between the ability to detect a genome that is truly present (measured using recall) and the incorrect detection of genomes not truly present (measured using precision). We therefore compare the performance of these methods, as well as simpler

methods based on aligned read count and breadth of coverage-based thresholds, using precision/recall curves (PRC) and further investigate their sensitivity to incomplete reference databases, their recall performance at low abundance, and their precision in the face of incomplete host removal. Finally, we present case-studies of the application of these methods to the profiling of mycobacterial communities present in a water distribution system[10] and *Neisseriaceae* present in the human oral cavity[14,25].

Materials Methods

Reference genome databases & simulated data sets

We assembled a reference genome database for the family *Mycobacteriaceae* (*n*=252) and used that of Chenal et al.[26] for *Neisseriaceae* (*n*=70), both using existing sequence data (NCBI genome accessions in Tables S1 and S2). These databases were curated to select the highest quality genome for each species, preserving named species even if they shared significant genomic similarity but have important phenotypic distinctions (e.g. *Neisseria meningitidis* and *N. gonorrhoeae*, ANI = 95% ; *Mycobacterium tuberculosis* and *M. canettii*, ANI = 99.2%). We also retained named subspecies of significance for *Mycobacteriaceae* (e.g. *M. abscessus* subsp *bolletii*, *massiliense*, and *abscessus,* ANI ~97%). For *Mycobacteriaceae*, a core genome phylogeny was estimated by first using roary[27] to produce a core genome alignment using the NCBI GFF3 gene annotation files (roary parameters -e mafft -i 60 -cd 99), followed by phylogenetic reconstruction using iqtree v2.2.2.7[28]. To illustrate genome relatedness, pairwise average nucleotide identity (ANI) was calculated using fastANI[29]. The same procedure was performed for the *Neisseriaceae* dataset using fastANI and the phylogeny estimated by Chenal et al.[26].

We then simulated 500 replicate shotgun metagenomes for each database (simulation procedure summarized in Fig. S1), consisting of Illumina sequencing reads simulated from a randomly selected subset of species present in our reference databases (variable *n* species per

replicate, drawn from normal distributions with means of 15 and 30, respectively and a standard deviation of 4). To represent the intraspecific genomic variability present in natural species: for each named species' reference genome, we randomly retrieved a corresponding genome sequence from the same named species from NCBI (RefSeq, all genomes assigned to *Mycobacteriaceae or Neisseriaceae*, retrieved using NCBI datasets, 2025/02/11, Supplementary Tables S3, and S4). Using a custom script (Supplementary materials, Zenodo archive, using the software packages fastANI 1.32[29], seqkit v.2.3.1[30], and samtools v.1.17[31], ChatGPT o4-mini-high was used as a coding aid to refine and troubleshoot scripts used in this manuscript, but not to conceive or create them), we verified the taxonomic identity of the chosen genome by comparing its ANI to our reference genome database and ensuring the reference genome for the (sub)species being simulated was the top ANI hit to the randomly selected genome, and had ANI > 95%. In the case where no suitable alternative genome was available for a given taxon or it was an unnamed species, the reference genome was used for simulations. We then simulated reads using Mason[32] (read length, 100bp) to achieve a mean depth drawn from an exponential distribution with a mean randomly selected from values between 0.01 and 10x (Fig. S1), injecting both simulated sequencing error and single nucleotide polymorphism (SNPs, at a rate randomly selected from 0.02, 0.01, 0.001, or 0.0001 substitutions per site), all while assuring an ANI of no less than 95% from the reference genome for the selected species was maintained in the final mutated genome used to simulate sequencing reads. Finally, simulated reads from each selected species were combined into a synthetic metagenome.

Execution of YACHT, metapresence, and sylph, breadth and read-based methods

To add context to the results, we also defined simple threshold-based species-level genome identification tools using read alignments. Reads were first competitively aligned to the respective concatenated reference database using bowtie2 (options –very-sensitive --end-to-

end)[33], and then filtered for alignment mapping quality (none, MAPQ2, 3, 5, 10, 20 and 30) using samtools. Read count and bread of coverage-based methods were implemented using scripts to process the results of samtools' coverage tool, and report a species present if the aligned read count or breadth of genomic (defined as fraction of target bases covered by a least one aligned read[6,34]) exceeded a predefined threshold. Metapresence[6] is an alignment-based method, so aligned reads contained in a BAM file, at varying mapping quality filter levels, were used to execute the metapresence script. YACHT[8] and sylph[5] were executed after building the respective reference k-mer sketches using default parameters. For each method, thresholds were varied among the relevant ranges. For metapresence, the thresholding parameters of the Fraction of Unexpected Gaps (FUG) and Breadth-Expected breadth Ratio (BER)[6] were varied along their ranges of 0-1. For YACHT, there are two main thresholding parameters: the ANI cutoff at which genomes are considered the same species, and the minimum exclusive breadth of coverage. Executions were performed at varying thresholds for each. Sylph can also be tuned for ANI threshold, which was allowed to vary between 0.95 and 0.99. Sylph has three other tuning parameters: -c, for the fraction of k-mers sampled for sketching of both genomes and reads, --min-count-correct, the k-mer multiplicity required for coverage correction, and --min-number-kmers which serve as a minimum threshold to output results. Our tests demonstrated essentially no meaningful impact of these parameters in this analysis, yielding the same mean precision and recall at ANI = 0.95 across replicates, so these were kept at the defaults.

Precision-recall curves

The relative performance of these methods to recover the genomes that were used to generate the simulated datasets was compared using precision-recall curves (PRC) after calculating the relevant statistics (true positive, false positives, false negatives, and true negatives) across the relevant parameter/threshold spaces for each method in each replicate

metagenome. Results are plotted for each threshold in precision/recall space at the mean of each metric across replicates (± standard deviation) using a python script (Supplementary materials, Zenodo archive) and the following packages : numpy[35], pandas[36], Scikit-learn[37] and matplotlib[38]. PRC curves were compared using area under the curve measurements, and optimal parameters were determined by selecting the set of parameters that maximized the F_1-score.

Recall in relation to simulated genome depth, and abundance estimation

A further analysis was conducted to measure recall as a function of the simulated genome depth, binning genomes by depth across all simulated metagenomes for each reference dataset using an R script (Supplementary materials, Zenodo archive). The read count-based threshold method was not tested as it was similar to the breadth-based method but performed less well (see results). For the breadth-based threshold and metapresence, the best performing breadth threshold and BER/FUG parameters from the PRC analysis were used for each dataset and alignment filtering stringency (Table 1). This analysis was also repeated using the actual simulated read counts. Recall was calculated as mean recall in a given bin, using a bootstrap resampling procedure with 1000 replicates to provide a measure of significance (95% bootstrapped range of means). Accuracy of abundance estimation was compared between methods by comparing the true simulated abundance to the predicted abundance by computing the normalized L_1 (Manhattan distance) metric.

Recall and precision in relation to incomplete host removal and the presence of other microbial reads

To assess the robustness of these methods to incomplete removal of host reads, we first performed an experiment using reads simulated from the human genome (NCBI Genome accession GCF_000001405.40). Using mason[32], we simulated Illumina reads from the human

genome to an average depth of 1.0x. We repeated this 10 times then combined these reads with a randomly selected sample of 10 of our simulated metagenome replicates from the *Mycobacteriaceae* and *Neisseriaceae* databases (Fig. S1) and repeated the execution of the methods, comparing precision between the 10 simulated metagenomes, with and without additional human host reads.

To determine whether the inclusion of other microbial sequences not reflected in the reference genome impacts precision, we simulated reads from microbial genomes drawn from release 220 (r220) of the Genome Taxonomy Database (GTDB)[39]. We first removed genomes attributed to our families of interest in the GTDB taxonomy, and then randomly selected 100 genomes from the GTDB r220, simulated reads using mason, injecting both read and SNP errors, at a random depth, as above, but without resampling genomes from NCBI. We repeated this 100 times, and concatenated these synthetic GTDB metagenomes with random samples of 100 each of the *Mycobacteriaceae* and *Neisseriaceae* simulated metagenomes. We then compared precision of the methods between with, and without the additional microbial reads.

The degree to which the precision of each method is affected by an incomplete reference database was assessed using a database subsampling experiment. For each reference database, a set of subsampled reference databases was created by randomly removing an increasing fraction of reference genomes, in steps of 10% of the total reference database size, generating 20 replicates of the subsampled reference genome database for each level of genome removal using a python script (Supplementary materials). Then, these sets of subsampled databases were used with the various methods to estimate precision at each level of genome removal using a randomly selected sample of 20 simulated metagenomes per family from the initial 500 metagenomes for each family, such that the samples may contain reads from species no longer present in the subsampled reference database.

Case studies

To illustrate the best performing methods' performance on real-world datasets, we retrieved two shotgun metagenomic datasets. For *Mycobacteriaceae*, we downloaded short-read metagenomes (NCBI Bioproject PRJNA1081894) derived from an interrogation of a chloraminated wastewater system dataset[10], for which both culture and a MAG-based approach was employed in the original article. Adapters were trimmed from the raw reads using fastp v0.240[40] and aligned against the *Mycobacteriaceae* reference database using bowtie2[33]. Sylph and metapresence were executed as above, using optimized BER/FUG parameters (Table 1) for metapresence. For *Neisseriaceae*, we downloaded oral cavity metagenomes from the dorsum of tongue and supragingival plaque of healthy humans from Phase I of the human microbiome project (HMP, SRA sample accessions in Table S5)[25], ensuring samples for each site were from a different individuals. HMP Phase I reads already have host (human) reads removed, and sequencing adapters trimmed. These reads were aligned to the *Neisseriaceae* database using bowtie2, with sylph and metapresence executed as above. For both datasets, we extracted the presence/absence and relative overall abundance from the outputs of sylph (using the -u option) and metapresence (obtaining absolute read count from the metapresence output), and scaling the abundances using the corresponding genome length and overall read count to obtain read per kilobase-million (RPKM). For illustrative purposes and comparability, as genome breadth is not directly reported by sylph, we calculated genome breadth for all genomes in the respective reference databases using the unfiltered bowtie2 read alignment and the samtools coverage tool. For illustrative purposes, we randomly selected 4 samples from the dorsum of the tongue, and 4 from supragingival plaque to plot in detail.

Results

Datasets

We first assembled reference genome databases for *Mycobacteriaceae* and *Neisseriaceae*. The assembled reference genome databases demonstrated a range of genome

relatedness (Fig. S2, and S3). The *Mycobacteriaceae* database contained several clades of named species sharing high ANI (e.g. *M. ulcerans*, *M. marinum*, and related species). The core genome phylogeny calculated here (Fig. S2) was similar to recently published analyses[41–43] and we have opted to use the traditional nomenclature of the family, with a single genus, *Mycobacterium*[44]. The *Neisseriaceae* reference database was notable both for the relatively high ANI within certain clades (e.g. *N. meningitidis*, *N. gonorrhoeae* and other related species of nasopharyngeal *Neisseria*, pair-wise ANI ~94-95%) and for significant horizontal gene leading to unexpectedly high ANI at odds with the core genome phylogeny (Fig. S3). For example, although relatively basal among *Neisseria* in the core genome tree, the human oral commensal *N. elongata* had relatively higher ANI relative to other human oral commensal *Neisseria*, suggesting HGT between these co-occurring species[15]. The overall mean±SD pairwise ANI in these family-level reference genome databases was 0.80±2.23 *Mycobacteriacaeae* and 0.81±4.62 for *Neisseriaceae*. The mean±SD number of genomes per replicate was 29.2±3.7 and 14.8±3.8, respectively for the *Mycobacteriaceae* and *Neisseriaceae* datasets. Median and interquartile range (IQR) for depth of coverage, simulated read count and breadth of genomic coverage (fraction of genome covered by at least one read[6,34]) across replicate simulated metagenomes was 0.23x (1.21), 6547 reads (34368), and 20% (68) for *Mycobacteriaceae* and 0.22x (1.17), 2539 reads (13928), and 19% (66) for *Neisseriaceae*. The mean final ANI between the selected, mutated genomes used to simulate reads (Fig. S1) and their corresponding reference genomes was 98.7±1.1 and 98.1±1.5, respectively.

Precision and Recall performance

We next assessed the performance of the five methods in order to examine their explicit and implicit trade-offs between recall (the ability of a given method to detect the presence of a given species that is truly present, also called true positive rate, sensitivity or completeness) and precision (the fraction of species identified as present by the method that are truly present in the

metagenomic sample, also called purity). These trade-offs were examined visually using precision/recall curves (PRCs, Fig. 1, Fig. S4,S5), and by comparing area-under-the curve (AUC) values for each method (Table 1). The simple breadth-based threshold with alignment filtering was the most optimal classifier tested by AUC criteria (Table 1), and performed better than read-count based thresholds (Fig. 1a,b,d,e). PRCs also demonstrated the importance of filtering aligned reads by mapping quality, which yielded better overall classifier performance for aligned read and breadth threshold-based metrics (Fig. 1a,b,d,e, Table 1). However, the alignment threshold-based approaches (breadth, read count) began to sacrifice precision if the threshold was set to achieve recall greater than 0.6-0.7 in these datasets. YACHT performed less well at lower ANI thresholds (Fig. 1c,f, Table 1). Metapresence performed well across both datasets (Fig. 1g,h,j,k), and provided very good recall before sacrificing precision, particularly when the input alignment is filtered for mapping quality (Fig. 1h,k, Fig. S4,S5). Sylph delivered moderate to good recall at very high precision, making very few false positive identifications (Fig. 1i,l). Sylph's tunable ANI threshold performs best when left at the default of 95% in these datasets. Results were largely congruent between the *Mycobacteriaceae* and *Neisseriaceae* datasets, although overall classifier performance, as measured by AUC, was lower for *Neisseriaceae*. This likely reflected both the higher overall higher degree of HGT and genome relatedness in this family and the lower absolute number of reads simulated (as simulations were parameterized based on a specified depth, and *Neisseriaceae* have shorter genomes, mean length of 2.4Mb vs 5.7Mb for *Mycobacteriaceae*). PRC curves, and the F_1-score were used to define optimal thresholding for each dataset (*Neisseriaceae*, *Mycobacteriaceae*) and method for further testing of recall and precision, below (Table 1).

Method	**Alignment Quality Filter**	***Mycobacteriaceae***	***Neisseriaceae***

		AUC	Threshold (%)		F_1-score	AUC	Threshold (%)		F_1-score
Breadth of genomic coverage threshold	none	0.869	2.0		0.794	0.855	3.3		0.749
	MAPQ3	0.954	0.6		0.885	0.903	2.1		0.808
	MAPQ20	0.971	0.3		0.906	0.955	0.5		0.878
			Threshold (n)		**F_1-score**		**Threshold (n)**		**F_1-score**
Read count threshold	none	0.819	1800		0.742	0.815	800		0.723
	MAPQ3	0.924	500		0.847	0.873	500		0.785
	MAPQ20	0.949	200		0.874	0.930	200		0.844
			Threshold		**F_1-score**		**Threshold**		**F_1-score**
			ANI	**coverage**			**ANI**	**coverage**	
YACTH	-	0.934	0.9999	.006	0.865	0.898	0.9999	0.009	0.802
			Threshold				**Threshold**		
Metapresence			**BER**	**FUG**	**F_1-score**		**BER**	**FUG**	**F_1-score**
	none	0.913	0.8	0.5	0.891	0.889	0.8	0.5	0.865
	MAPQ3	0.966	0.5	0.5	0.952	0.920	0.6	0.5	0.900
	MAPQ20	0.965	0.2	0.3	0.942	0.940	0.3	0.4	0.926
			Threshold (ANI)		**F_1-score**		**Threshold (ANI)**		**F_1-score**
Sylph	-	0.833	0.95		0.802	0.825	0.95		0.789

Table 1. Performance of methods to identify genomes present in a metagenomic sample based on 500 simulated metagenomic datasets of species of *Mycobacteriaceae* and

Neisseriaceae. Optimized threshold parameters are determined using the F_1-score. For alignment-based methods, results for mapping quality filtering at MAPQ3 and MAPQ20 are shown.

Recall performance in relation to depth and abundance and accuracy of relative abundance estimates

For all methods, overall recall improved with increasing simulated dataset depth and absolute read abundance, reaching close to perfect recall above depth > 0.1x for all methods (Fig. 2a,b,d,e). Recall for the breadth-based threshold method varied with alignment quality filtering, related to the optimized threshold chosen (Table 1). Notably, sylph had lower recall at low read abundance, with poor recall at depths of coverage less than 0.1x, which likely explains the lower maximal recall observed in the P/R curves, as our simulated datasets had low overall median depth (0.22-0.23x). Metapresence had a much lower limit of detection, and was able to identify the presence of genomes from as little as hundreds of reads (Fig 2d,e). YACHT's recall performance at low abundance is intermediate between sylph and metapresence at the optimized thresholds (Fig 2a,b,d,e, Table 1). The recall performance of metapresence demonstrated somewhat unexpected dynamics related to alignment filtering and thresholding parameters (BER/FUG), as the algorithm switches between consideration of both BER and FUG to only BER at higher read depths[6]. Further, at very high read depth > 2.0x, the expected BER against which the BER is compared becomes stricter while calculated BER saturates, which likely explains the decrease in recall observed at high simulated depth. When executed with default rather than optimized BER and FUG thresholds, these dynamics are exaggerated (Fig. S6). With stricter alignment filtering (MAPQ20), metapresence achieved slightly lower recall at higher abundance relative to other methods, particularly with default thresholds (Fig. S6). Also unexpectedly, moderate alignment quality (MAPQ3) filtering increased rather than decreased recall at low abundance (Fig 2.a,c, Fig. S6), which highlights the complex interplay between

alignment mapping quality filtering, BER/FUG thresholds, and read spatial distribution: moderately strict filtering decreases overall aligned read count, but in the case of a genome truly present, may make the read distribution more even, thus increasing calculated BER and FUG[6].

The accuracy of each method to recapitulate the original relative abundance of the individual species in the replicate megenomes was compared using the normalized L_1 distance (Fig 2c,f; Table S6). YACHT does not directly report relative abundance, and so was not included. Sylph and the breadth of coverage threshold method were the most accurate methods, with a mean normalized L_1 of 0.14 for both the *Mycobacteriaceae* and *Neisseriaceae* datasets, while metapresence was slightly less accurate with 0.20 and 0.29, respectively. For all alignment-based methods there was a clear trend toward less accurate abundance estimates at higher alignment quality filtering levels, particularly at MAPQ20 (Table S6).

Precision performance in the presence of other genomes and incomplete host read removal

Execution of the methods with non-removed host reads revealed little impact on the precision (Fig. 3a,d; Table S7). Further, inclusion or not of reads from additional microbial genomes simulated at a varying depth and error likewise had little impact on precision (Fig. 4b,e, Table S7). When metapresence is executed without alignment quality filtering, small differences are observed in precision between the two datasets, with precision seemingly increasing (from median precision 0.87 to 0.91) when the additional microbial reads from outside the families are included in the *Neisseriaceae* dataset (Fig. 3e). This may be related to homology among these additional reads aligning poorly with the genomes of interest, thus increasing the total read count, and thus affect the BER and FUG calculation. When alignment quality filtering was used with metapresence, or with the sylph method, there is no appreciable impact on the precision of the methods: thus these methods are generally robust to the presence of reads from other microbial genomes or the human genomes at the thresholds used.

Precision performance in the setting of an incomplete reference database

Precision of the breadth threshold method decreased markedly with reference database subsampling (Fig 3c,f), with alignment quality filtering increasing precision overall but not altering the trajectory of precision loss with a sparse reference genome database. The *Mycobacteriaceae* dataset showed higher overall precision, which may relate to the larger size of the reference database. YACHT demonstrated similar dynamics (Fig. 3g,j). Metapresence and sylph show greater robustness of precision to reference genome database subsampling (Fig. 3h,i,k,l), with metapresence was affected to a greater degree than sylph (Fig. 3g,h,i,l). Sylph is relatively robust to the removal of up to approximately 50% of the reference genome sequences before a drop in mean precision below 0.9 was observed, but variation in the range of precision estimates obtained at high subsampling fractions included 1.0 (Fig 3i,l). Metapresence demonstrated decreased precision after 20-30% of reference genome removal. Unexpectedly, as the subsampling fraction increased, more stringent alignment quality filtering (MAPQ) had a greater relative decay in precision in both datasets (Fig 3. h,k), which likely results from an interaction between the BER/FUG thresholds optimized using the full database and the alignment quality filtering and is less marked with the default, unoptimized BER/FUG thresholds are used (Fig. S7). Confidence intervals around estimates are wide, likely reflecting an effect of sampling bias at very low subsampling fractions.

Case studies

We reanalyzed the Dowdell et al.[10] dataset of shotgun metagenomic reads characterizing the mycobacterial community in a chloraminated water distribution system (Fig. 4, S8) and examined the distribution of *Neisseriaceae* in the human oral cavity using sequencing data from the human microbiome project[25] (Fig. 5, S9, S10, S11). We elected to use

the two methods, sylph and metapresence/MAPQ20, with the best overall performance in terms of recall, precision and robustness to reference database subsampling (see also discussion).

Mycobacterial communities in a drinking water system

This analysis demonstrates the greater overall sensitivity of these reference-based approaches relative to the MAG approach adopted by Dowdell et al. Both sylph (Fig. 4) and metapresence/MAPQ20 (Fig. S8) identify substantially more genomes, including those identified by pooled culture-based methods, but not by MAG (e.g. *M. franklinii*, *M. aurum*, and *M. chelonae*). As expected, genomes with high breadth of genomic coverage are identifiable in both MAG and reference-based approaches, although *M. gadium* for example is present at > 50% breadth of coverage in the Site B / December sample and was not recovered by Dowdell et al.. Despite the high theoretical sensitivity of culture for mycobacteria, in practice cultivability depends on many factors, and standard approaches may not be optimal for specific mycobacteria[9]. As predicted by the in silico simulations, at low abundance, metapresence with MAPQ20 alignment mapping filter identifies (Fig. 4) the presence of more low abundance species relative to sylph (Fig S8). Dowdell et al. also identified a mycobacterial MAG that they were not able to identify, but this sequence is not available to compare with or augment our reference database, and so may not be reflected in this analysis.

Presence of specific *Neisseriaceae* in different oral metagenomes

Analysis of the tropism of *Neisseriaceae* in oral metagenomes demonstrates differences in abundance and niche within even the small confines of the human oral cavity (randomly selected subset of 8 samples using metapresence/MAPQ20 in Fig. 5, and sylph in Fig. S9, complete HMP Phase I data with metapresence/MAPQ20 and sylph in Figs S10,S11). *Neisseria* known to inhabit the oral cavity and the tongue are primarily identified from the dorsum of the tongue (e.g. *N. subflava*), while genera and species associated with teeth/gingiva are recovered

mostly, and at higher abundance, from the supragingival plaque samples (e.g. *N. elongata*, *Eikenella spp.*, *Kingella spp.*). These results are in general congruent with those of Donati et al.[14] and Albanese and Donati[45] who used these same HMP data, but employed strain-based approaches: respectively, reference-based alignment followed by SNP calling, as well as MLST profiles[14] and StrainEst[45]. The high degree of genomic relatedness illustrates the value of the sylph and metapresence methods to increase precision of species identification, as many species traditionally associated with the nasopharynx, and thus likely not truly present (e.g. *N. meningitidis*) have significant breadth of genomic coverage by read alignment, but are not identified as present by metapresence or sylph, except for a single sample (Fig 5, Fig. S9, S10, S11). There small difference between the genomes identified by these methods, particularly when adjudicating the presence of very closely genomes present at high abundance (e.g. *N. maccacae* and *N. sicca*, pairwise ANI=96.5), and in general metapresence/MAPQ identifies more species as present (mean 6.6 species per metagenome) relative to sylph (5.7).

Discussion

Methods to determine the presence of a given genome in a metagenome balance the ability to detect a genome (recall/true positive rate/sensitivity/completeness) against the rate of false positive identifications (precision/purity)[23]. In this use case for these methods, maintaining a high precision is important - false positive genome identifications are generally to be avoided as it is more conservative to conclude there is lack of evidence to support the presence of a genome than to conclude the incorrect genome is present. By analogy to alternative approaches to identify species of interest - specific PCR or culture are generally relatively precise but sometimes not sensitive (such as when the bacteria is fragile or difficult to grow). However, the ideal balance between recall and precision may depend on the exact hypothesis being tested and underscores the importance of a clear understanding of the tradeoffs implicit in each method. Fundamentally, there are two sources of information considered by these methods: 1)

the degree to which the reads diverge from the reference sequence (ANI, alignment and mapping quality) and 2) the spatial distribution of reads and overall breadth of coverage of reads relative to the reference genome. The datasets studied here, of closely related species in groups with horizontal gene transfers[24,46,47] and simulated at low overall abundances, incorporating natural genomic variation, likely represent a challenging use case for these algorithms, where consideration of both sources of information is important, as demonstrated by the overall better performance of the metapresence algorithm and breadth of coverage threshold if the input alignment is filtered for mapping quality (Fig. 1).

Overall, this simulation study suggests that the best performing methods for the identification of genomes at the species-level (defined pragmatically), at low abundance and among closely related taxa, are metapresence and sylph, the later of which has previously been demonstrated to be superior for species-level profiling to older methods like kraken2/bracken[5]. We note that it is important that read alignments are filtered for mapping quality for metapresence (Fig. 1, Table 1), although this has the effect of distorting relative abundance measures (Fig 2,c,f). Therefore, caution is urged in interpreting relative abundance in filtered alignments, which may be used as an initial presence/absence filter prior to recalculation of abundance using an unfiltered alignment. Simpler alignment threshold-based methods using breadth of coverage and especially read count lack precision in these datasets, particularly if the reference genome database is incomplete (Fig. 3), and enforce thresholds which are dataset specific (Table 1). YACHT performed relatively less well in these datasets, which may be related to its ANI thresholding parameter - whereby genomes with pairwise ANI above this threshold are collapsed during sketching[8], which is particularly problematic for the *Neisseriaceae* database (Fig. S3). Sylph offers high precision and acceptable overall recall, with the lower recall observed here explained by the relatively higher lower limit of detection (here approximately depth 0.1x, Fig. 2a,d) in these datasets characterized by relatively low overall depth. Sylph also accurately recovers the original relative abundances simulated, and is also

faster than most others methods[5], and so would be particularly suited to large scale analyses of thousands of metagenomes, particularly at higher absolute sequencing depths. Sylph is also more robust to an incomplete genome database, retaining a reasonable degree of precision even in the absence of up to 50% missing genome sequences from the reference database (Fig. 3). However, the declining precision of these methods in the face of an incomplete reference genome database urges caution in metagenomic studies where a high proportion of uncharacterized taxa or “microbial dark matter” may be present[48], for example in environmental samples, like soils[9]. For example, an *hsp65* PCR-based study of mycobacterial diversity in global soils found a large majority (> 90%) of amplified sequences were not attributable to known species[9]. This finding suggests a hybrid approach where including as many unique MAGs as possible in the reference database would likely increase the precision of species-level metagenomic profiling, while retaining the added recall at low abundance afforded by a reference-based approach[5]. These conclusions are based on in silico simulations that attempt to capture dynamics of natural genomic variability that lead to genomic divergence of reads from the reference genome within the bounds of real, rather than ideal, species definitions. A limitation of this in silico approach is that dynamics related to relative efficacy of DNA extraction and sequencing between species, among other sampling and wet-lab factors which may bias precision and recall, were not captured.

These results are difficult to compare with those obtained by the authors of the studies describing these methods[5,6,8], as the methods compared by each author, the nature of datasets used, and simulation procedures all differed somewhat. Among the methods examined here, only metapresence and YACHT have been directly compared[6]. In that study of a mock community generated using an expanded CAMISIM reference genome dataset[49], the optimal parameters for YACHT were found to be similar as in this study (ANI = 0.99, min. coverage = 0.05)[6]. Metapresence was found to have higher balanced accuracy and recall (true positive rate) relative to YACHT, particularly when the mock community was created with highly unequal

abundance, which is consistent with the greater recall seen here at low sequence abundance (Fig 2a,b,d,e). Our estimate of sylph's recall at low abundance and precision are quite similar to those reported by that tool's authors[5], where a depth of coverage greater than ~0.1x is necessary before genomes are reliably detected (Fig 2a,d) and that precision is generally very high (> 0.9) across a range of conditions (Fig 1h,k, Fig. 3a,b,d,e,i,l).

These approaches should facilitate the study of species-level presence and abundance of a targeted group of species. In the first of two real-world metagenomic datasets examined here, the results of both sylph (Fig. 4) and metapresence (Fig. S8) show that this approach identified more genomes than a MAG-based approach. Although this study of mycobacterial communities in a drinking water system did not include an exact gold standard as culture results were pooled, and identification by mass spectroscopy may be imprecise[12], sylph and metapresence detected several species found in culture but not by the MAG approach (e.g. *M. chelonae, M. aurum*). This dataset also illustrates some advantages of culture-independent techniques for the profiling of mycobacterial communities[12], which captured several species (in the case *M. aurpense*, convincingly so), not recovered in culture for a given site. Overall, a more diverse mycobacterial community is suggested by the sylph and metapresence analyses relative to either the MAG or culture-based analysis, including the identification of biologically plausible species recovered from other water systems like *M. gadium*[50,51] and *M. frederiksbergense*[52] missed by the other approaches. In the second real-world example, the human oral cavity is home to a complex microbiome, where multiple *Neisseriaceae* species not only occupy overlapping niches, but also readily exchange genetic material[24]. The increasing availability of high quality genomic reference sequences in combination with the methods tested here allow straightforward and rapid delineation of the tropism of *Neisseriaceae* in the oral cavity (Fig. 5, S9, S10, S11), recapitulating the main patterns of tropism seen in strain level methods[14,45].

The results of this study demonstrate the feasibility of accurate species-level identification of a defined group of interest within metagenomic samples using these state-of-the-art algorithms. We show that sylph and metapresence (with alignment filtering) can be used for profiling known species in a specific taxonomic group and are robust, to varying degrees, to both incomplete host-sequence removal, and the absence of genomic reference sequences for species outside of the group of interest. Despite their promise, these methods have limitations that shape their ideal use cases. Metapresence with alignment quality filtering has high recall, but suffers from an important loss of precision in the face of a highly incomplete reference. For these reasons, metapresence may be well suited to human microbiome and related studies - where high recall in the face of low biomass is advantageous, and the majority of species likely present have been characterized at the genomic level (either by MAG or traditional approaches). For sylph, recall is limited at very low abundance, an important consideration for pathogen detection and low-biomass studies but is balanced by very high precision that is maintained despite a significantly incomplete reference database. This combination of features and the algorithm's speed would be most advantageous in the face of a highly incomplete reference database and high biomass, as would be expected in, for example in deeply sequenced soil samples.

Acknowledgments

L.B.H received funding from the Fonds de Recherche du Québec - Santé/Québec Ministère de la Santé et des Services sociaux Clinician-Scientist Training Program. F.J.V. received a Junior 1, Junior 2, and Senior research scholar salary award from the Fonds de Recherche du Québec - Santé. This research was enabled in part by support provided by Calcul Québec and the Digital Research Alliance of Canada (alliancecan.ca). This work was supported by the Natural

Sciences and Engineering Research Council of Canada (NSERC) discovery grant (RGPIN-2023-05657) and Canadian Institute of Health Research (CIHR) project grant (450862). Funders had no role in the conception or conduct of the research reported in this manuscript. The authors thank Dr. Marcel Behr for a critical reading of an early draft of this manuscript.

Data/Code availability

Scripts used to generate the synthetic metagenomes, process the results, and generate the figures are available in a Zenodo repository (10.5281/zenodo.15678063, https://doi.org/10.5281/zenodo.15678063). Truthsets and the results for each method are also included in the repository. All genome information used in this analysis are publicly available, with genome accession numbers found in Tables S1, S2, S3, S4, and metagenomic reads for the Dowdell et al. dataset available on NCBI, under Bioproject PRJNA1081894 and those from the Human Microbiome Project also available on NCBI, with SRA accession numbers in Table S5.

Author Contributions

L.B.H. Conceived the analysis, performed experiments, wrote the initial draft, reviewed and revised the manuscript.

J.O.A. Performed experiments, reviewed and revised the manuscript.

F.J.V. Conceived the analysis, financial/logistical support, reviewed and revised the manuscript.

References

1. Varma-Basil, M. *et al.* Development of a Novel PCR Restriction Analysis of the hsp65 Gene as a Rapid Method To Screen for the Mycobacterium tuberculosis Complex and Nontuberculous Mycobacteria in High-Burden Countries. *J. Clin. Microbiol.* **51**, 1165–1170

(2013).

2. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).

3. Nayfach, S. *et al.* A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* **39**, 499–509 (2021).

4. Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504 (2019).

5. Shaw, J. & Yu, Y. W. Rapid species-level metagenome profiling and containment estimation with sylph. *Nat. Biotechnol.* 1–12 (2024) doi:10.1038/s41587-024-02412-y.

6. Sanguineti, D., Zampieri, G., Treu, L. & Campanaro, S. Metapresence: a tool for accurate species detection in metagenomics based on the genome-wide distribution of mapping reads. *mSystems* **9**, e00213-24 (2024).

7. Shen, W. *et al.* KMCP: accurate metagenomic profiling of both prokaryotic and viral populations by pseudo-mapping. *Bioinformatics* **39**, (2023).

8. Koslicki, D., White, S., Ma, C. & Novikov, A. YACHT: an ANI-based statistical test to detect microbial presence/absence in a metagenomic sample. *Bioinformatics* **40**, btae047 (2024).

9. Walsh, C. M., Gebert, M. J., Delgado-Baquerizo, M., Maestre, F. T. & Fierer, N. A Global Survey of Mycobacterial Diversity in Soil. *Appl. Environ. Microbiol.* **85**, e01180-19 (2019).

10. Dowdell, K. S. *et al.* Source-to-tap investigation of the occurrence of nontuberculous mycobacteria in a full-scale chloraminated drinking water system. *Appl. Environ. Microbiol.* **90**, e00609-24 (2024).

11. Zhang, H. *et al.* Applications and advances in molecular diagnostics: revolutionizing non-tuberculous mycobacteria species and subspecies identification. *Front. Public Health* **12**, (2024).

12. Mercaldo, R. A. *et al.* Environmental risk of nontuberculous mycobacterial infection: Strategies for advancing methodology. *Tuberculosis* **139**, 102305 (2023).

13. Harrison, L. B. & Behr, M. A. Mutation rates and the molecular epidemiology of *Mycobacterium abscessus*. *Proc. Natl. Acad. Sci.* **120**, e2307499120 (2023).
14. Donati, C. *et al.* Uncovering oral Neisseria tropism and persistence using metagenomic sequencing. *Nat. Microbiol.* **1**, 16070 (2016).
15. Nyongesa, S. *et al.* Evolution of longitudinal division in multicellular bacteria of the Neisseriaceae family. *Nat. Commun.* **13**, 4853 (2022).
16. Honda, J. R. Environmental Sources and Transmission of Nontuberculous Mycobacteria. *Clin. Chest Med.* **44**, 661–674 (2023).
17. Hakim, D. *et al.* Zebra: Static and Dynamic Genome Cover Thresholds with Overlapping References. *mSystems* **7**, e00758-22 (2022).
18. Peabody, M. A., Van Rossum, T., Lo, R. & Brinkman, F. S. L. Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. *BMC Bioinformatics* **16**, 1–19 (2015).
19. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial “pan-genome”. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 13950–13955 (2005).
20. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
21. Gihawi, A. *et al.* Major data analysis errors invalidate cancer microbiome findings. *mBio* **14**, e01607-23 (2023).
22. Blanco-Míguez, A. *et al.* Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol.* **41**, 1633–1644 (2023).
23. Meyer, F. *et al.* Critical Assessment of Metagenome Interpretation: the second round of challenges. *Nat. Methods* **19**, 429–440 (2022).
24. Marri, P. R. *et al.* Genome Sequencing Reveals Widespread Virulence Gene Exchange among Human Neisseria Species. *PLOS ONE* **5**, e11835 (2010).

25. Gevers, D. *et al.* The Human Microbiome Project: A Community Resource for the Healthy Human Microbiome. *PLOS Biol.* **10**, e1001377 (2012).
26. Chenal, M. *et al.* Discovery of the widespread site-specific single-stranded nuclease family Ssn. *Nat. Commun.* **16**, 2388 (2025).
27. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).
28. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
29. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).
30. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation | PLOS One. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0163962.
31. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
32. Holtgrewe, M. Mason: a read simulator for second generation sequencing data. (2010) doi:10.17169/refubium-22374.
33. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
34. Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* **15**, 121–132 (2014).
35. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).
36. McKinney, W. Data Structures for Statistical Computing in Python. *scipy* (2010) doi:10.25080/Majora-92bf1922-00a.
37. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Mach. Learn. PYTHON*.
38. Bisong, E. Matplotlib and Seaborn. in *Building Machine Learning and Deep Learning*

Models on Google Cloud Platform: A Comprehensive Guide for Beginners (ed. Bisong, E.) 151–165 (Apress, Berkeley, CA, 2019). doi:10.1007/978-1-4842-4470-8_12.

39. Parks, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **50**, D785–D794 (2022).
40. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
41. Behra, P. R. K. *et al.* Comparative genome analysis of mycobacteria focusing on tRNA and non-coding RNA. *BMC Genomics* **23**, 704 (2022).
42. Gupta, R. S., Lo, B. & Son, J. Phylogenomics and Comparative Genomic Studies Robustly Support Division of the Genus Mycobacterium into an Emended Genus Mycobacterium and Four Novel Genera. *Front. Microbiol.* **9**, (2018).
43. Val-Calvo, J. & Vázquez-Boland, J. A. Mycobacteriales taxonomy using network analysis-aided, context-uniform phylogenomic approach for non-subjective genus demarcation. *mBio* **14**, e02207-23 (2023).
44. Meehan, C. J., Barco, R. A., Loh, Y. H. E., Cogneau, S. & Rigouts, L. Reconstituting the genus mycobacterium. *Int. J. Syst. Evol. Microbiol.* **71**, (2021).
45. Albanese, D. & Donati, C. Strain profiling and epidemiology of bacterial species from metagenomic sequencing. *Nat. Commun.* **8**, 2260 (2017).
46. Veyrier, F., Pletzer, D., Turenne, C. & Behr, M. a. Phylogenetic detection of horizontal gene transfer during the step-wise genesis of Mycobacterium tuberculosis. *BMC Evol. Biol.* **9**, 196 (2009).
47. Reva, O., Korotetskiy, I. & Ilin, A. Role of the horizontal gene exchange in evolution of pathogenic Mycobacteria. *BMC Evol. Biol.* **15**, 1–8 (2015).
48. Ma, B. *et al.* A genomic catalogue of soil microbiomes boosts mining of biodiversity and genetic resources. *Nat. Commun.* **14**, 7318 (2023).

49. Fritz, A. *et al.* CAMISIM: simulating metagenomes and microbial communities. *Microbiome* **7**, 17 (2019).
50. Virdi, R. *et al.* Lower Recovery of Nontuberculous Mycobacteria from Outdoor Hawai'i Environmental Water Biofilms Compared to Indoor Samples. *Microorganisms* **9**, 224 (2021).
51. Moghaddam, S. *et al.* Isolation of nontuberculous mycobacteria species from different water sources: a study of six hospitals in Tehran, Iran. *BMC Microbiol.* **22**, 1–9 (2022).
52. Loret, J.-F. & Dumoutier, N. Non-tuberculous mycobacteria in drinking water systems: A review of prevalence data and control means. *Int. J. Hyg. Environ. Health* **222**, 628–634 (2019).

Figures

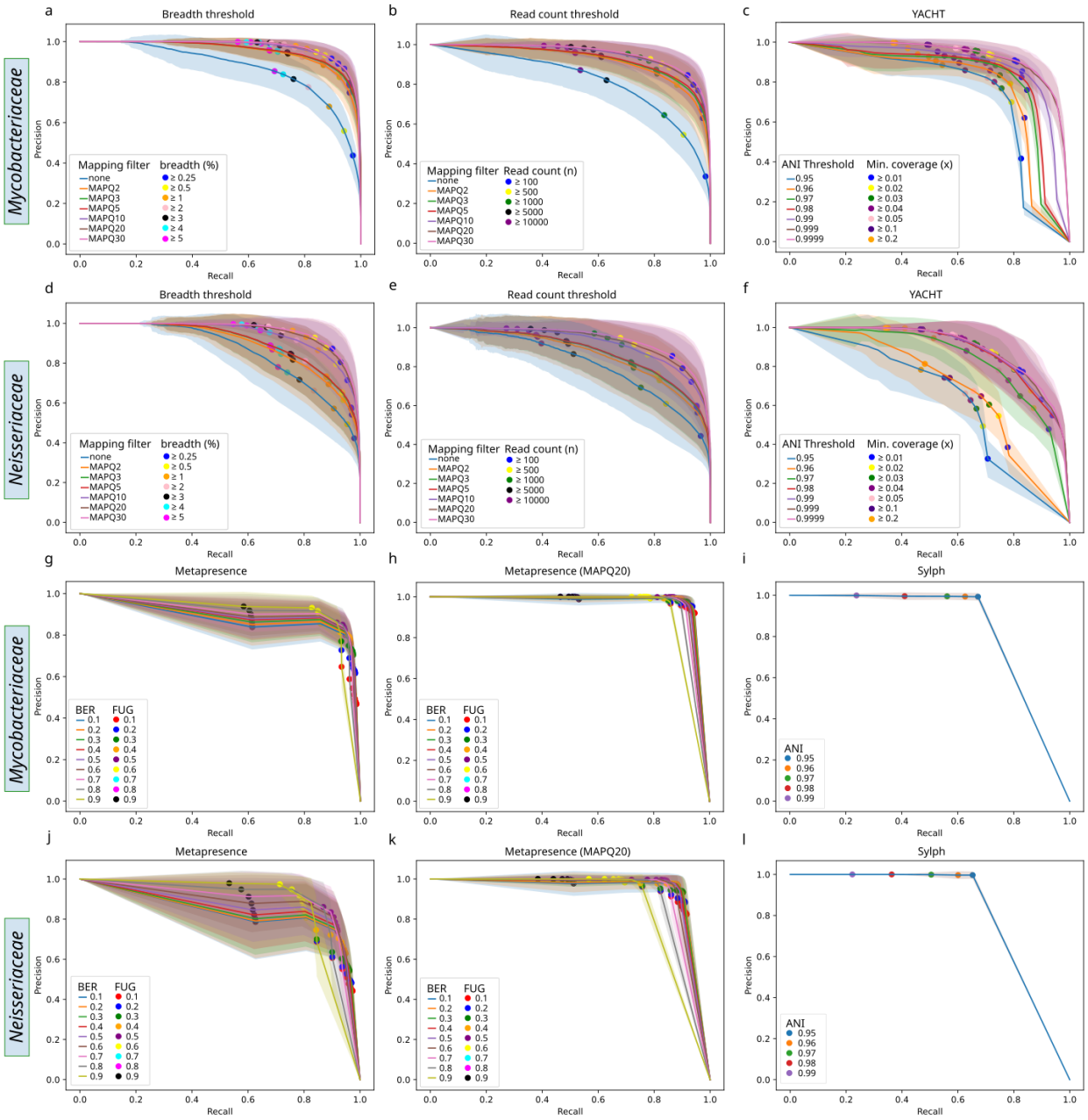


Figure 1. Precision-recall curves (PRCs) for each method, calculated across 500 simulated metagenomes for each of *Mycobacteriaceae* (a-c, j-l) and *Neisseriaceae* (d-f,g-i). Each panel illustrates the trade-off between precision and recall for each method, PRCs are plotted at the mean precision and recall across replicates, with the standard deviation across replicates shaded. Each panel includes a sample of specific thresholds used to define the curves. Visualized methods are simple aligned breadth of coverage (a,d) and

read count (b,e) threshold-based methods, both unfiltered and filtered for mapping quality (MAPQ2,3,4,5,10,20,30); YACHT (c,f) executed using five different average nucleotide identity (ANI) thresholds, varying the threshold of minimum coverage; Metapresence, executed on the unfiltered read alignment (g,j). Fraction of Unexpected Gaps (FUG) and Breadth-Expected breadth Ratio (BER) refer to thresholding parameters used by metapresence. Metapresence, executed on the read alignment, filtered for mapping quality at MAPQ20 (h,k). Sylph (i,j), executed with default parameters, varying the ANI parameter.

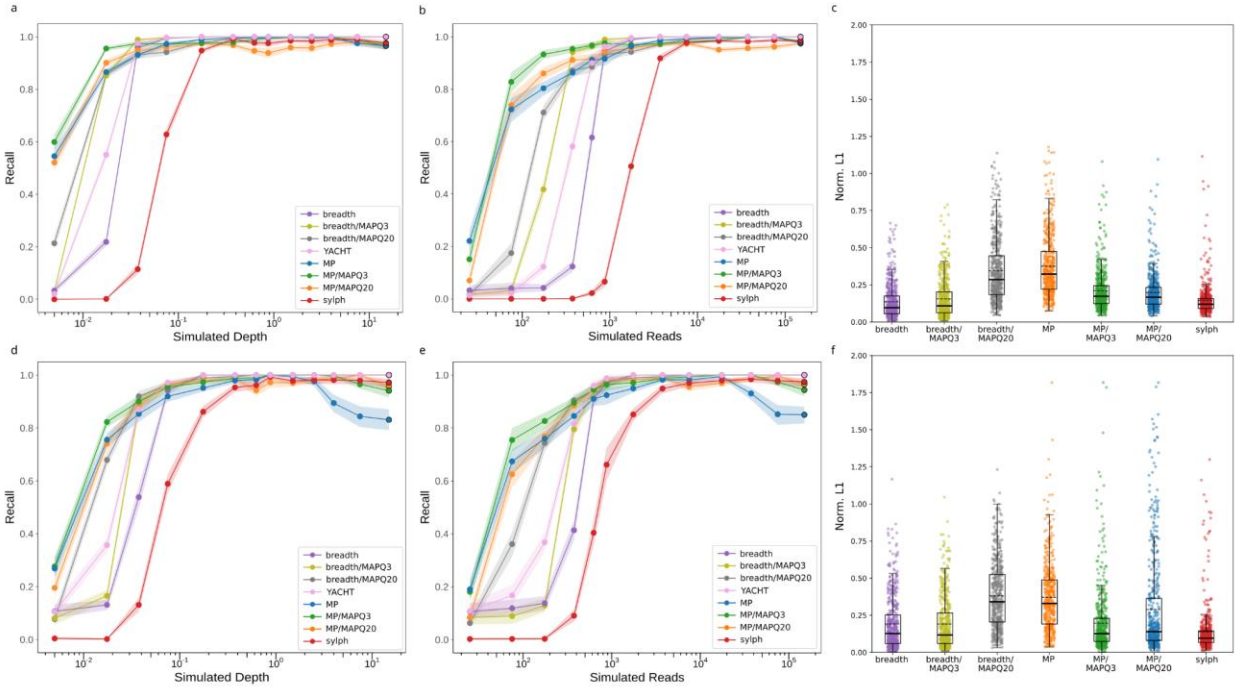


Figure 2. Recall plotted against binned read depth (a,d) used to simulate metagenomic reads as well as the actual number of simulated reads generated (b,e). The x-axis is log-transformed and mean recall is given at the bin midpoint. Results are depicted for the (a,b) *Mycobacteriaceae* and (c,d) *Neisseriaceae* datasets. Recall is summarized in each depth bin (bin boundaries: 0, 0.01x, 0.025x, 0.05x, 0.1x, 0.25x, 0.5x, 0.75x, 1.0x, 2.0x, 3.0x, 5.0x, 10.0x, 10.0x-max) and read count bin (bin boundaries: 0, 50, 100, 250, 750, 1000, 2500, 5000, 10000, 25000, 50000, 100000, 100000-max) across replicates of the simulated metagenomes on a per genome and method basis. The mean recall for the last bin in each plot is placed at an arbitrary x value within the last bin (as the upper border is infinite) and is plotted with a solid dark circle outline. Confidence intervals around each method's estimated recall were computed by a bootstrap resampling procedure using 1000 replicates. Accuracy of relative abundance estimates of the different methods are summarized in (e,f) using the normalized L_1 metric, with median and interquartile range depicted using box plots, and the mean normalized L1 shown at a dashed line.

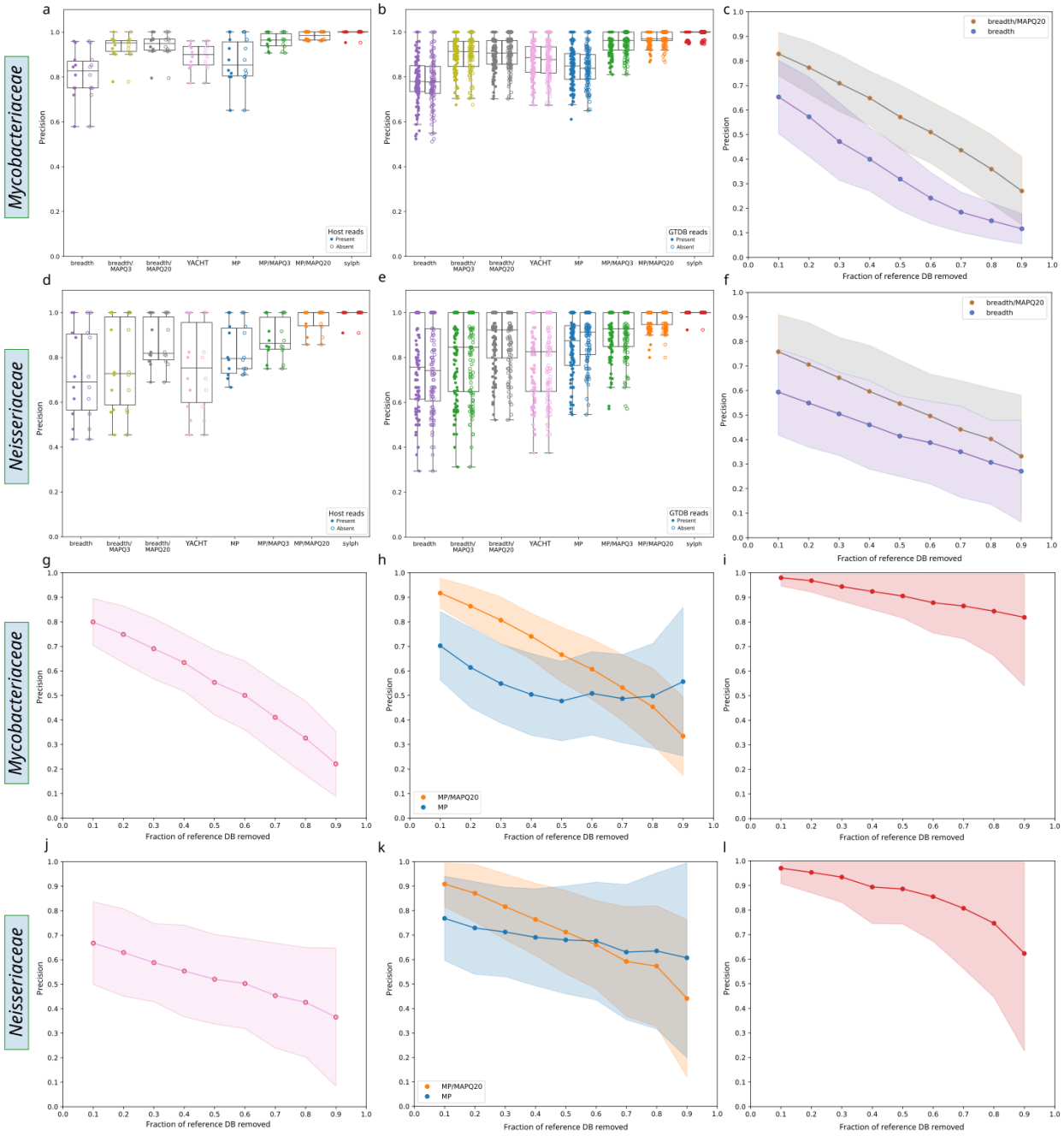


Figure 3. Precision of the examined methods in relation to the presence of reads outside the group of interest (a,b,d,e) and as a function of reference genome database subsampling (c,f,g-l), across the two datasets examined: *Mycobacteriaceae* (a-c, g-i) and *Neisseriaceae* (d-f, j-l). Precision the face of incomplete host removal (a,d) is calculated both with and without genomic reads from the human genome. Precision in the face of other microbial reads (b,e) is calculated

both with and without genomic reads derived from 100 randomly sampled genomes from the Genome Taxonomy (GTDB) database. In both cases, data points represent individual replicates, and boxplots median and interquartile range. Performance of the methods in the face of an incomplete reference database is shown by subsampling the reference database and recalculating precision at each level of subsampling for the breadth based-classifier (c,f), YACHT (g,j), metapresence (h,k) and sylph (i,l). For the subsampling analysis, points are plotted at mean precision across replicates, and the shaded area represents the standard deviation of precision across replicates.

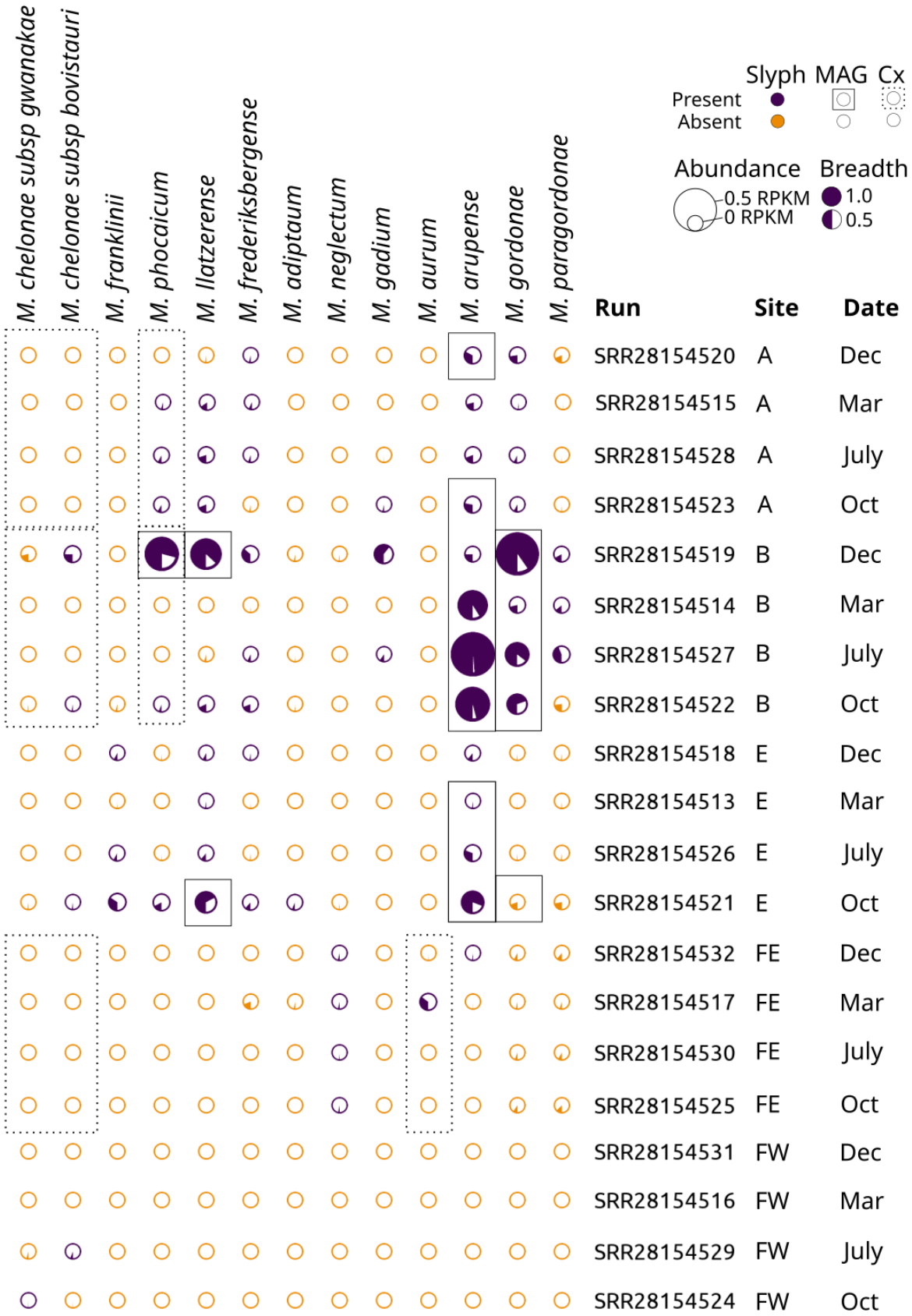

M. chelonae subsp gwanakae
M. chelonae subsp bovistauri
M. franklinii
M. phocaicum
M. llatzerense
M. frederiksbergense
M. adiptaum
M. neglectum
M. gadium
M. aurum
M. arupense
M. gordonae
M. paragordonae
Run
Site
Date
SRR28154520 A Dec
SRR28154515 A Mar
SRR28154528 A July
SRR28154523 A Oct
SRR28154519 B Dec
SRR28154514 B Mar
SRR28154527 B July
SRR28154522 B Oct
SRR28154518 E Dec
SRR28154513 E Mar
SRR28154526 E July
SRR28154521 E Oct
SRR28154532 FE Dec
SRR28154517 FE Mar
SRR28154530 FE July
SRR28154525 FE Oct
SRR28154531 FW Dec
SRR28154516 FW Mar
SRR28154529 FW July
SRR28154524 FW Oct
Slyph MAG Cx
Present
Absent
Abundance
0.5 RPKM
0 RPKM
Breadth
1.0
0.5

Figure 4. Presence/absence, abundance, and breadth of genomic coverage of mycobacterial species in the metagenomes of a chloraminated drinking water system by the sylph, culture and meta-assembled genome (MAG) approaches. This plot relates genomes, ordered phylogenetically as columns, to their estimated presence and abundance in the metagenomes of Dowdell et al[10]. Only mycobacterial species for which any method has been identified are shown. The small circles are shaded according to the breadth of genomic coverage based on read alignments and are scaled in size by the abundance in reads per kilobase-million (RPKM), based on the relative abundance reported by slyph (-u option) and scaled by the total read count. Circles colored in purple are identified as present by the sylph method. Those identified by the MAG approach used by Dowdell et al., are identified by a solid square around the circular breadth/abundance plots, while those identified by culture-based methods by Dowdell et al. (pooled by site) are identified by the dashed squares.

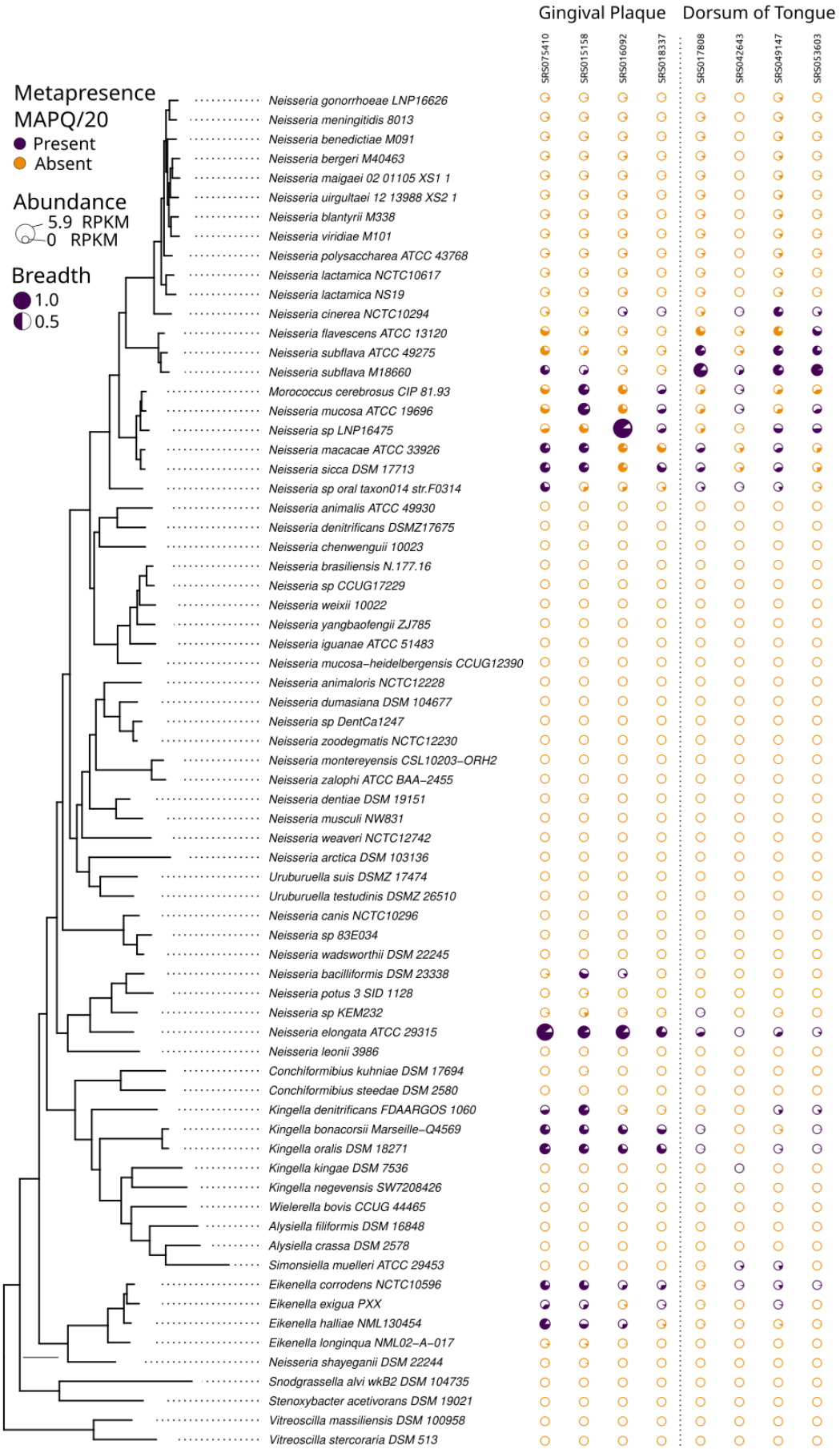
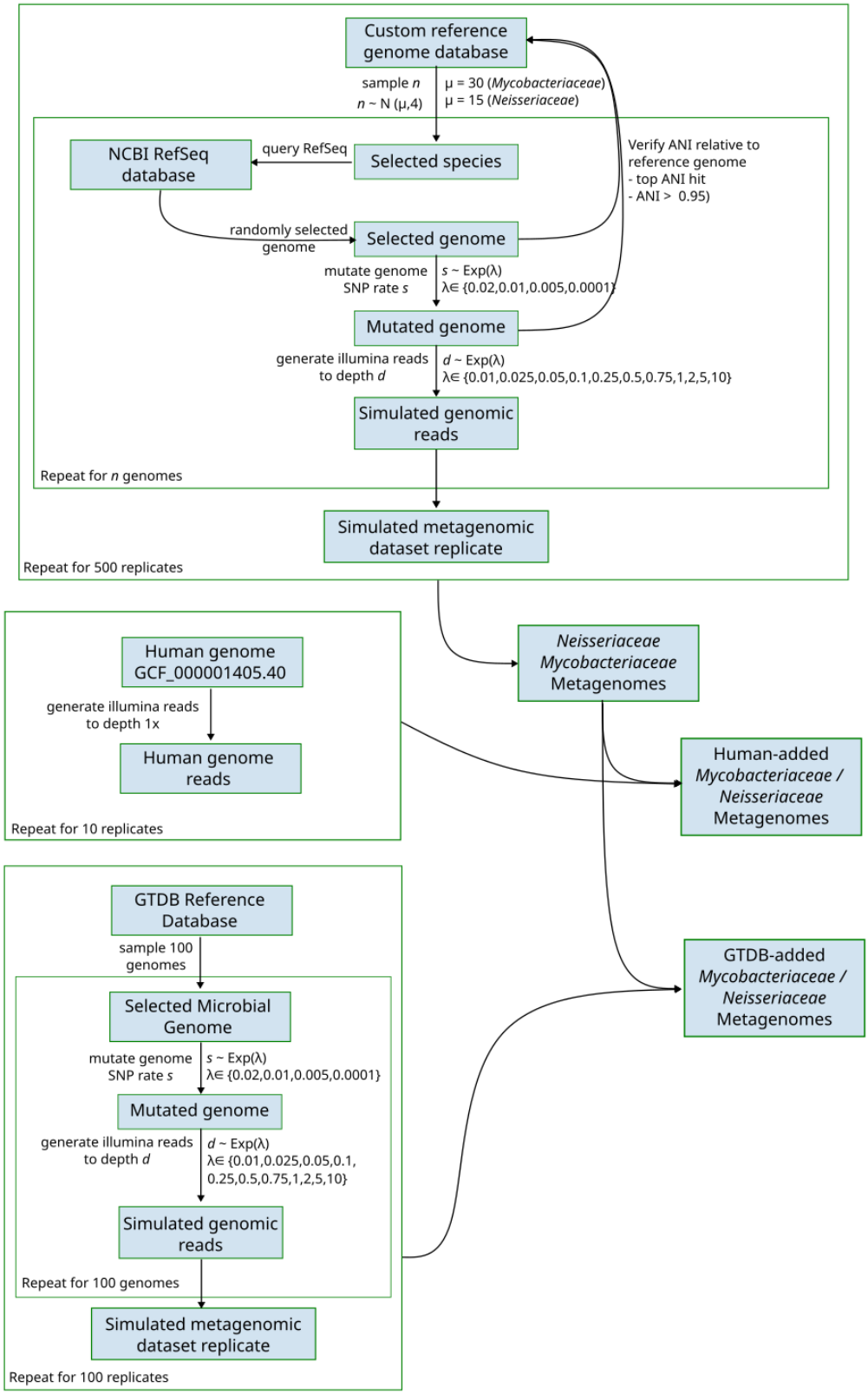
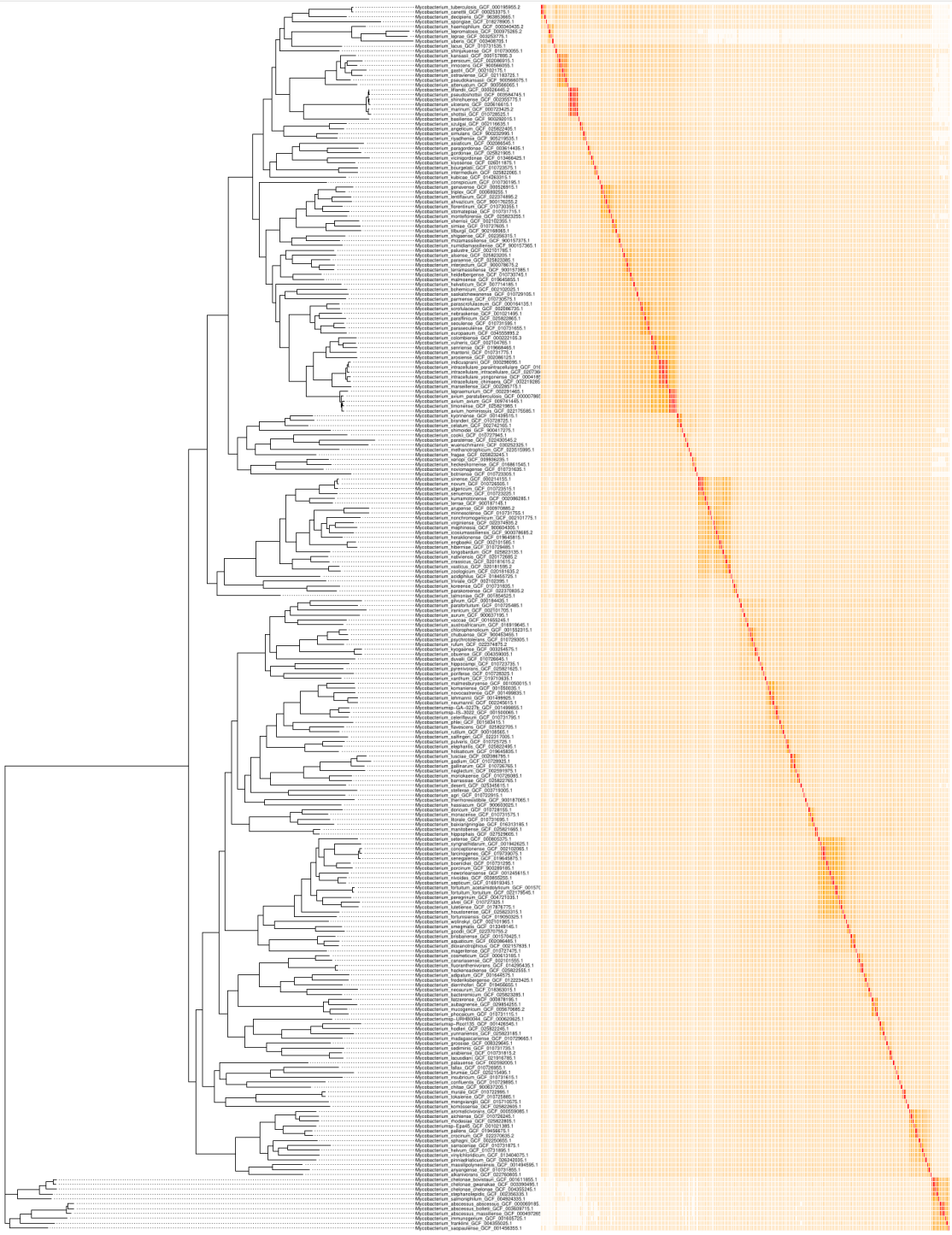

Gingival Plaque
Dorsum of Tongue
SRS075410
SRS015158
SRS016092
SRS018337
SRS017808
SRS042643
SRS049147
SRS053603
Metapresence
MAPQ/20
Present
Absent
Abundance
5.9 RPKM
0 RPKM
Breadth
1.0
0.5
Neisseria gonorrhoeae LNP16626
Neisseria meningitidis 8013
Neisseria benedictiae M091
Neisseria bergeri M40463
Neisseria maigaei 02 01105 XS1 1
Neisseria uirgultaei 12 13988 XS2 1
Neisseria blantyrii M338
Neisseria viridiae M101
Neisseria polysaccharea ATCC 43768
Neisseria lactamica NCTC10617
Neisseria lactamica NS19
Neisseria cinerea NCTC10294
Neisseria flavescens ATCC 13120
Neisseria subflava ATCC 49275
Neisseria subflava M18660
Morococcus cerebrosus CIP 81.93
Neisseria mucosa ATCC 19696
Neisseria sp LNP16475
Neisseria macacae ATCC 33926
Neisseria sicca DSM 17713
Neisseria sp oral taxon014 str.F0314
Neisseria animalis ATCC 49930
Neisseria denitrificans DSMZ17675
Neisseria chenwenguii 10023
Neisseria brasiliensis N.177.16
Neisseria sp CCUG17229
Neisseria weixii 10022
Neisseria yangbaofengii ZJ785
Neisseria iguanae ATCC 51483
Neisseria mucosa−heidelbergensis CCUG12390
Neisseria animaloris NCTC12228
Neisseria dumasiana DSM 104677
Neisseria sp DentCa1247
Neisseria zoodegmatis NCTC12230
Neisseria montereyensis CSL10203−ORH2
Neisseria zalophi ATCC BAA−2455
Neisseria dentiae DSM 19151
Neisseria musculi NW831
Neisseria weaveri NCTC12742
Neisseria arctica DSM 103136
Uruburuella suis DSMZ 17474
Uruburuella testudinis DSMZ 26510
Neisseria canis NCTC10296
Neisseria sp 83E034
Neisseria wadsworthii DSM 22245
Neisseria bacilliformis DSM 23338
Neisseria potus 3 SID 1128
Neisseria sp KEM232
Neisseria elongata ATCC 29315
Neisseria leonii 3986
Conchiformibius kuhniae DSM 17694
Conchiformibius steedae DSM 2580
Kingella denitrificans FDAARGOS 1060
Kingella bonacorsii Marseille−Q4569
Kingella oralis DSM 18271
Kingella kingae DSM 7536
Kingella negevensis SW7208426
Wielerella bovis CCUG 44465
Alysiella filiformis DSM 16848
Alysiella crassa DSM 2578
Simonsiella muelleri ATCC 29453
Eikenella corrodens NCTC10596
Eikenella exigua PXX
Eikenella halliae NML130454
Eikenella longinqua NML02−A−017
Neisseria shayeganii DSM 22244
Snodgrassella alvi wkB2 DSM 104735
Stenoxybacter acetivorans DSM 19021
Vitreoscilla massiliensis DSM 100958
Vitreoscilla stercoraria DSM 513

Figure 5. Presence/absence, abundance and breadth of genomic coverage of *Neisseriaceae* species in metagenomic samples from two sites in the oral cavities of humans according to the metapresence method with MAPQ20 alignment filtering. Eight samples from the Human Microbiome Project representing healthy individuals across two sites: the dorsum of the tongue and supragingival plaque. NCBI sequence read archive (SRA) sample accession numbers are provided in the figure for each metagenome. All *Neisseriaceae* species included in Chenal et al. are shown along with their phylogenetic relationships estimated by core genome analysis[26]. Circles are shaded according to the breadth of genomic coverage based on the unfiltered read alignments and are scaled in size by the abundance in reads per kilobase-million (RPKM), based on the filtered read alignments, as reported by metapresence. Circles colored in purple are identified as present by the metapresence tool.

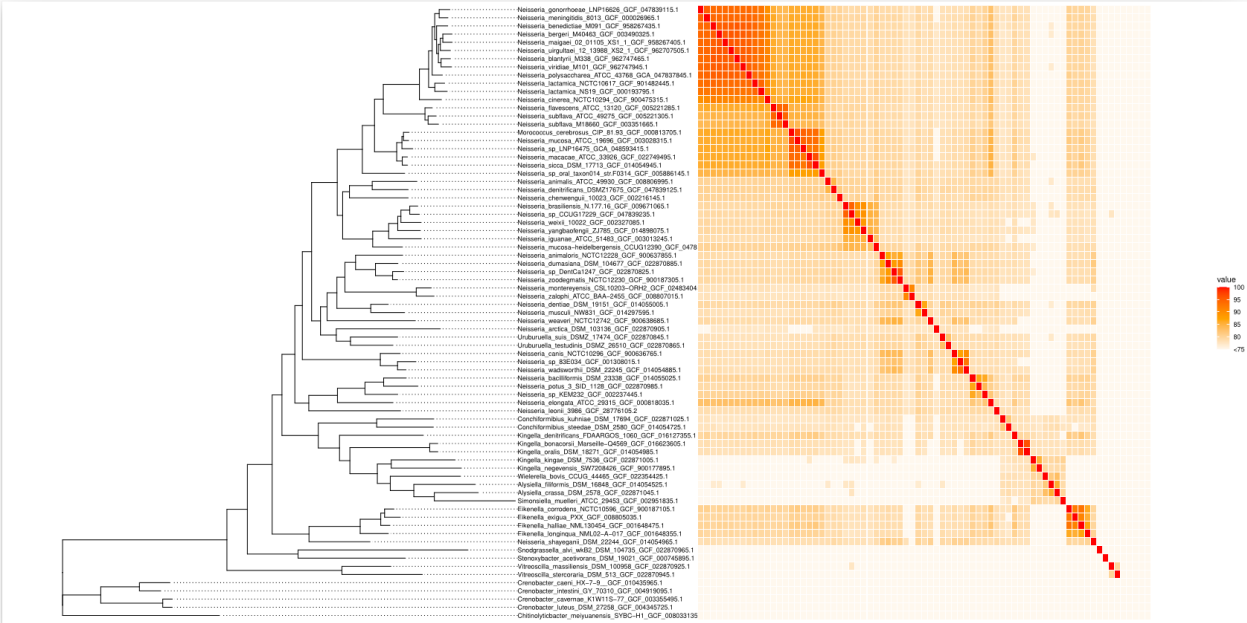
Supplementary Figures



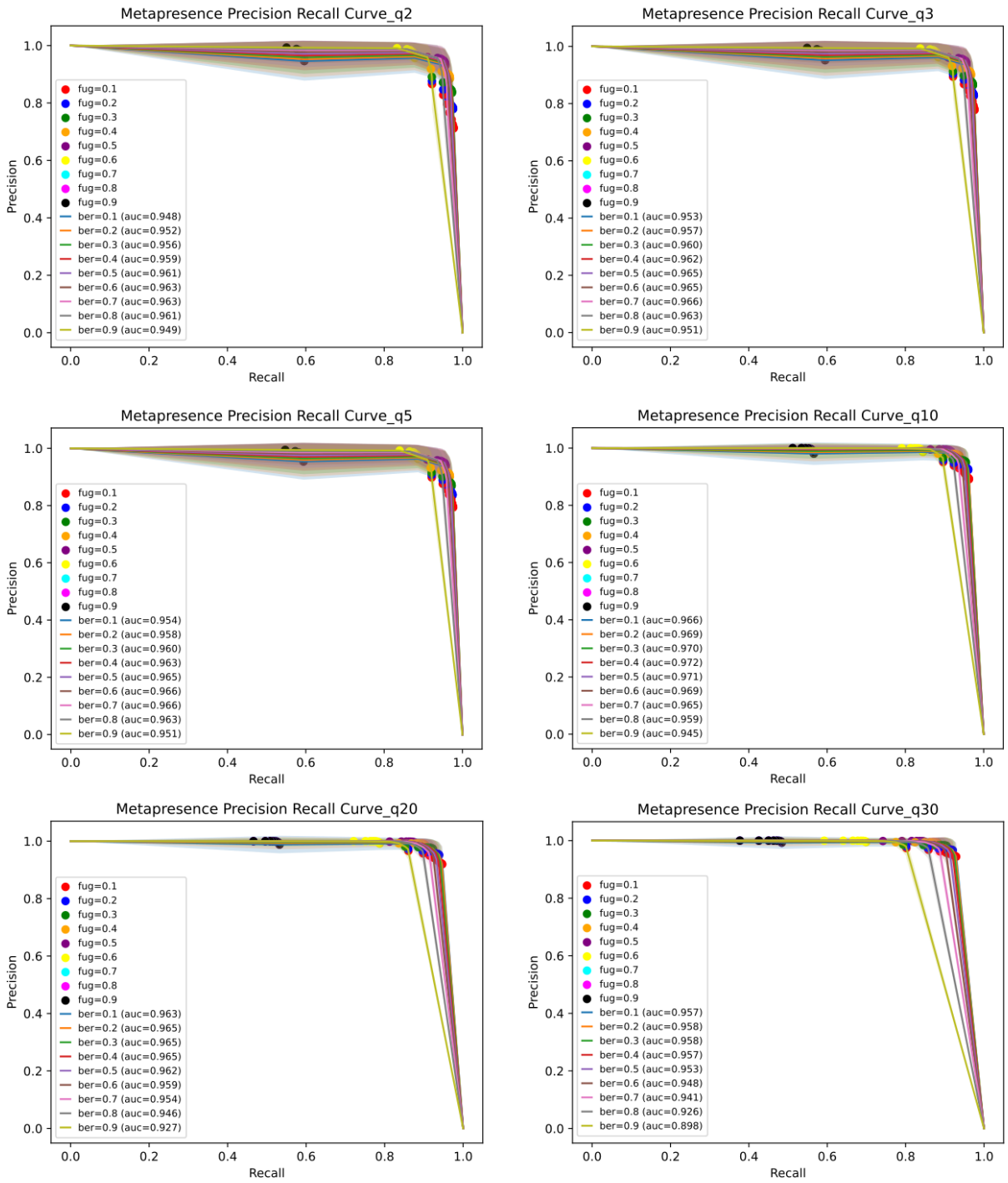
Supplementary Figure 1. Diagrammatic summary of the simulation procedure to generate synthetic metagenomes used in this study.


value
100
95
90
85
80
<75

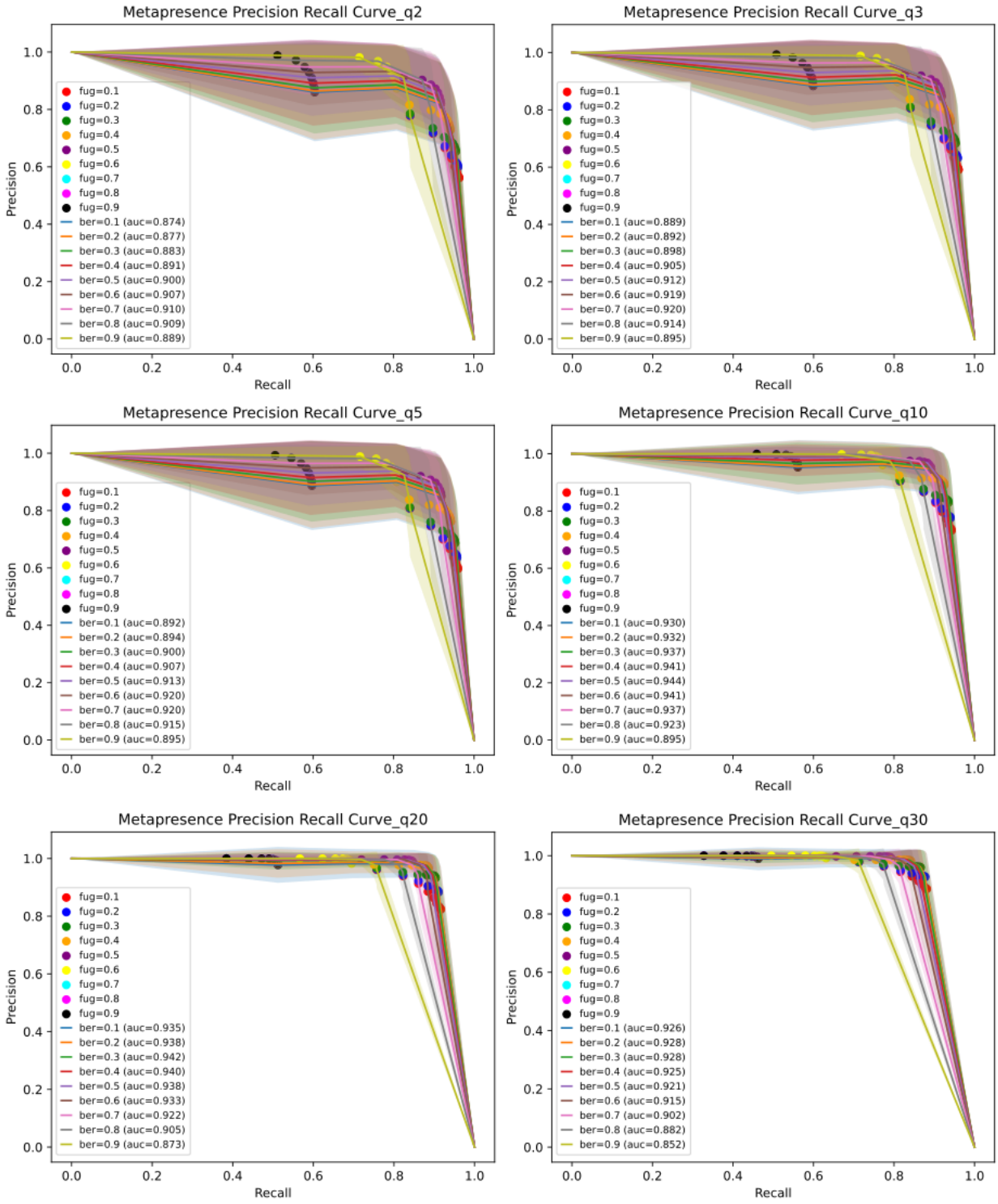
Supplementary Figure 2. Core genome phylogeny of the *Mycobacteriaceae* reference database used for read simulation. Pairwise average nucleotide identity (ANI) is illustrated on the right side of the plot.



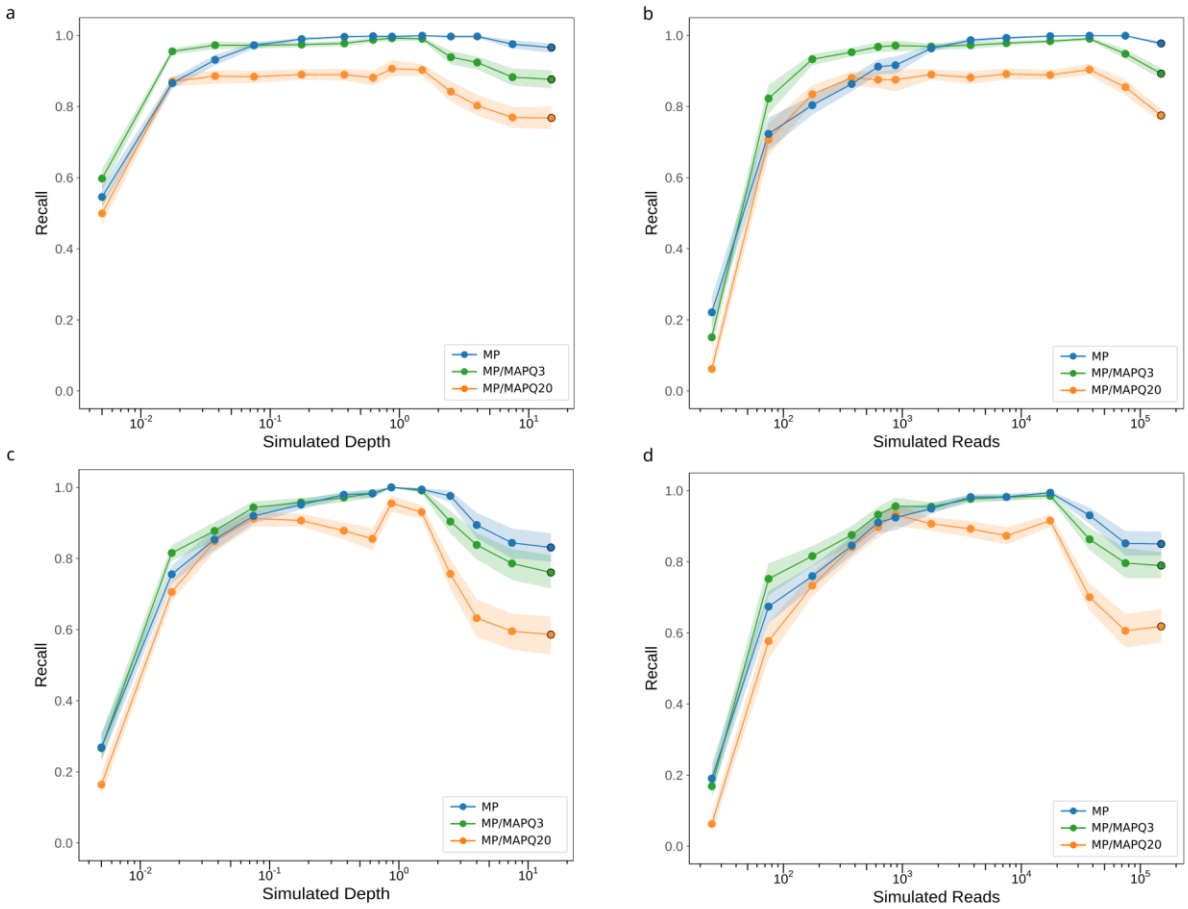
Supplementary Figure 3. Core genome phylogeny of the *Neisseriaceae* reference database used for read simulation. Pairwise average nucleotide identity (ANI) is illustrated on the right side of the plot. Several species of *Neisseriales* (*Crenobacter* sp. And *Chitiniloytibacter meiyuanensis*) are included as an outgroup for illustrative purposes but were not used in the analyses presented here.



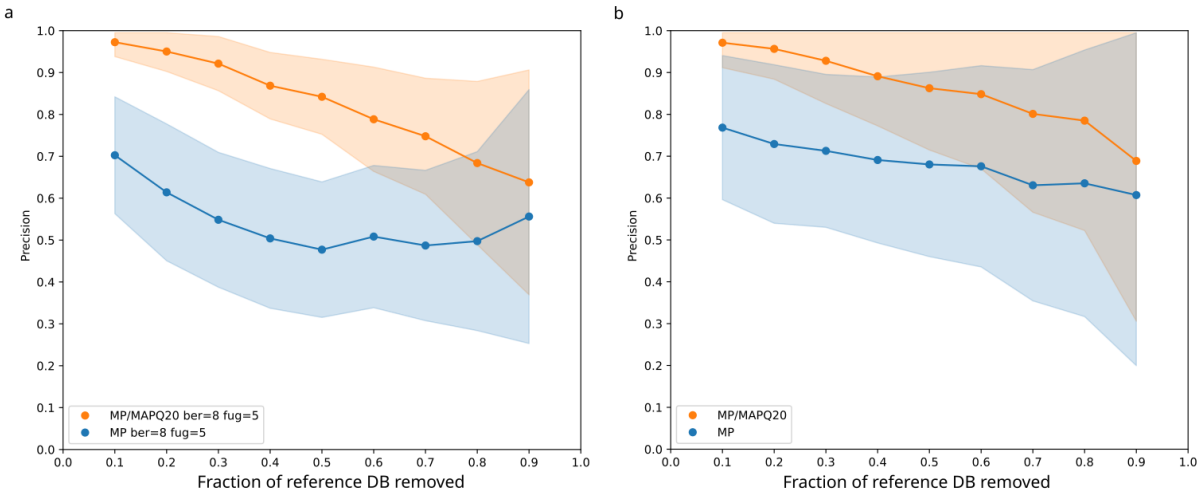
Supplementary Figure 4: Additional precision/recall curves for metapresence, at all alignment quality filtering levels (MAPQ 2,3,5,10, 20 and 30) for the *Mycobacteriaceae* dataset.



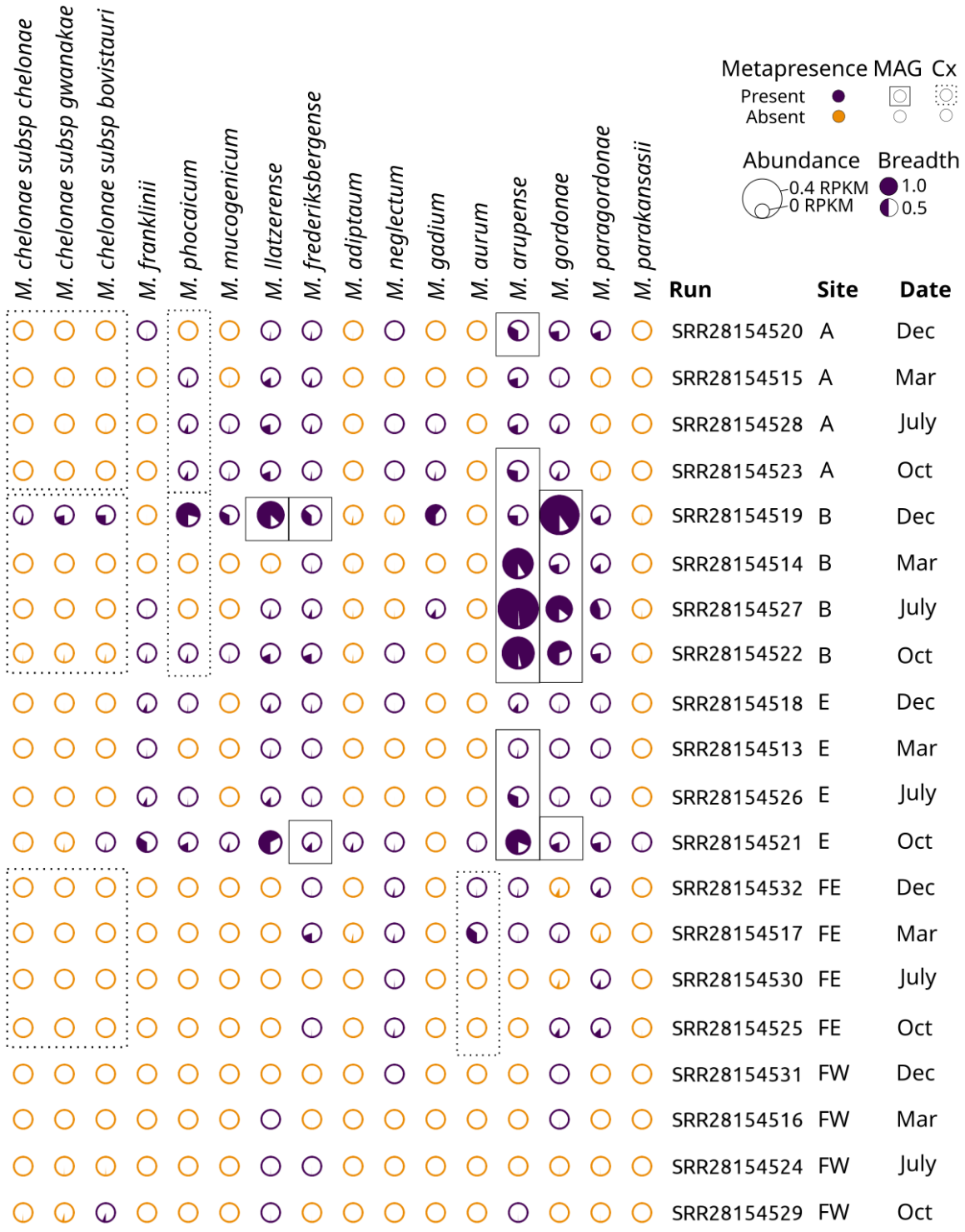
Supplementary Figure 5: Additional precision/recall curves for metapresence, at all alignment quality filtering levels (MAPQ 2,3,5,10, 20 and 30) for the *Neisseriaceae* dataset.



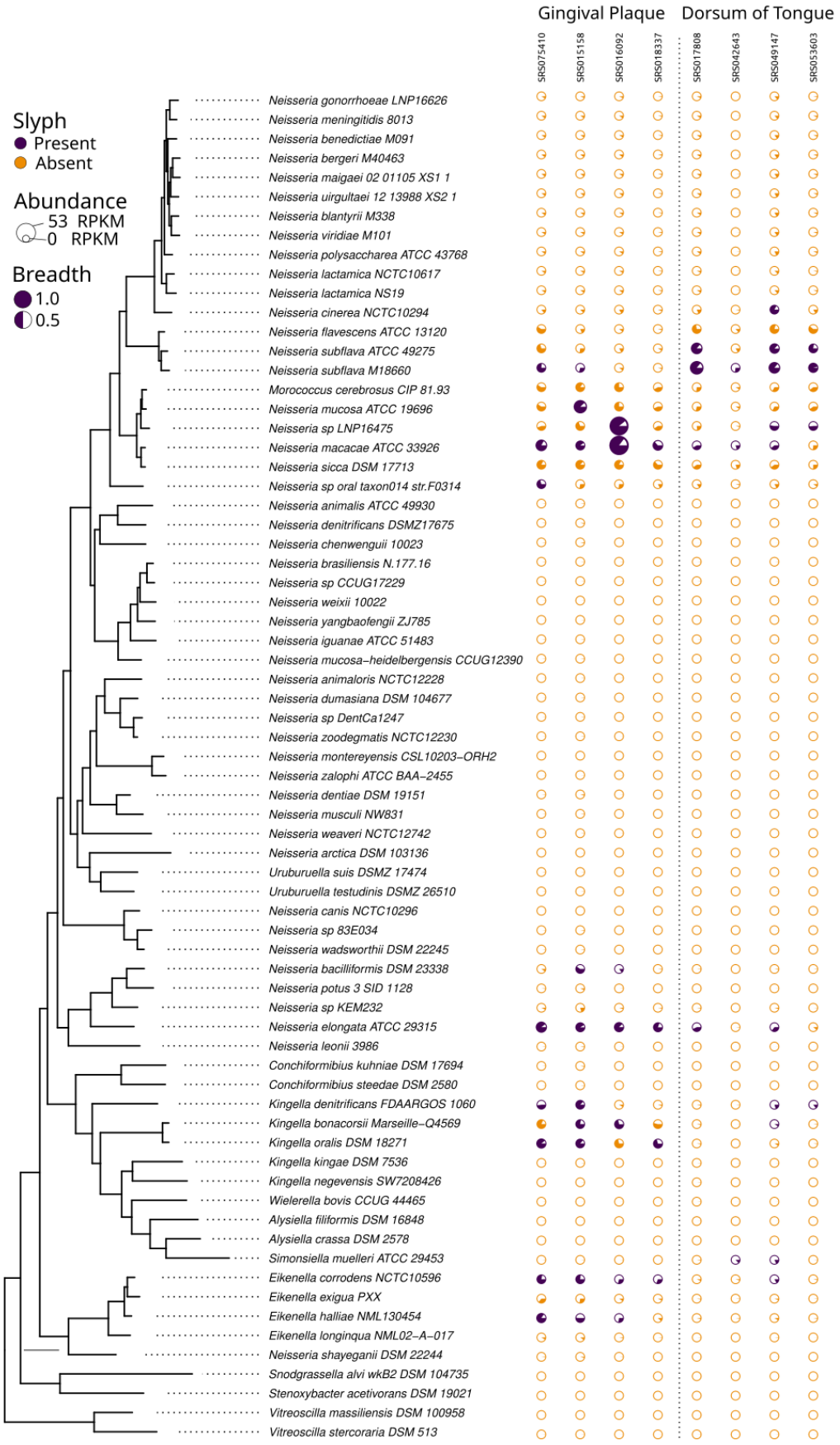
Supplementary Figure 6: Recall of the metapresence method as a function of simulated depth (a,c) and simulated reads (b,d) for the *Mycobacteriaceae* (a,b) and *Neisseriaceae* (c,d) simulated datasets. In these plots, metapresence was executed with default thresholds for BER and FUG: 0.8 and 0.5 respectively, rather than optimized values.



Supplementary Figure 7: Precision of the metapresence method plotted against subsampling of the reference genome databases (*Mycobacteriaceae*: a and *Neisseriaceae*: b) using default rather than optimized parameters for the BER and FUG thresholding parameters.


M. chelonae subsp chelonae
M. chelonae subsp gwanakae
M. chelonae subsp bovistauri
M. franklinii
M. phocaicum
M. mucogenicum
M. llatzerense
M. frederiksbergense
M. adiptaum
M. neglectum
M. gadium
M. aurum
M. arupense
M. gordonae
M. paragordonae
M. parakansasii
Metapresence
Present
Absent
MAG
Cx
Abundance
0.4 RPKM
0 RPKM
Breadth
1.0
0.5
Run
Site
Date
SRR28154520 A Dec
SRR28154515 A Mar
SRR28154528 A July
SRR28154523 A Oct
SRR28154519 B Dec
SRR28154514 B Mar
SRR28154527 B July
SRR28154522 B Oct
SRR28154518 E Dec
SRR28154513 E Mar
SRR28154526 E July
SRR28154521 E Oct
SRR28154532 FE Dec
SRR28154517 FE Mar
SRR28154530 FE July
SRR28154525 FE Oct
SRR28154531 FW Dec
SRR28154516 FW Mar
SRR28154524 FW July
SRR28154529 FW Oct

Supplementary Figure 8: Presence/absence, abundance, and breadth of genomic coverage of mycobacterial species in the metagenomes of a chloraminated drinking water system by the metapresence with MAPQ alignment filtering compared with different methodological approaches. This plot relates genomes, ordered phylogenetically as columns, to their estimated presence and abundance in the metagenomes of Dowdell et al[10]. Only mycobacterial species for which any method has been identified are shown. The small circles are shaded according to the breadth of genomic coverage based on read alignments and are scaled in size by the abundance in reads per kilobase-million (RPKM), based on the read count reported by metapresence/MAPQ20. Circles colored in purple are identified as present by the metapresence method. Those identified by the MAG approach used by Dowdell et al., are identified by a solid square, while those identified by culture-based methods by Dowdell et al. (pooled by site) are identified by the dashed squares.


Gingival Plaque
Dorsum of Tongue
SRS075410
SRS015158
SRS016092
SRS018337
SRS017808
SRS042643
SRS049147
SRS053603
Slyph
Present
Absent
Abundance
53 RPKM
0 RPKM
Breadth
1.0
0.5
Neisseria gonorrhoeae LNP16626
Neisseria meningitidis 8013
Neisseria benedictiae M091
Neisseria bergeri M40463
Neisseria maigaei 02 01105 XS1 1
Neisseria uirgultaei 12 13988 XS2 1
Neisseria blantyrii M338
Neisseria viridiae M101
Neisseria polysaccharea ATCC 43768
Neisseria lactamica NCTC10617
Neisseria lactamica NS19
Neisseria cinerea NCTC10294
Neisseria flavescens ATCC 13120
Neisseria subflava ATCC 49275
Neisseria subflava M18660
Morococcus cerebrosus CIP 81.93
Neisseria mucosa ATCC 19696
Neisseria sp LNP16475
Neisseria macacae ATCC 33926
Neisseria sicca DSM 17713
Neisseria sp oral taxon014 str.F0314
Neisseria animalis ATCC 49930
Neisseria denitrificans DSMZ17675
Neisseria chenwenguii 10023
Neisseria brasiliensis N.177.16
Neisseria sp CCUG17229
Neisseria weixii 10022
Neisseria yangbaofengii ZJ785
Neisseria iguanae ATCC 51483
Neisseria mucosa−heidelbergensis CCUG12390
Neisseria animaloris NCTC12228
Neisseria dumasiana DSM 104677
Neisseria sp DentCa1247
Neisseria zoodegmatis NCTC12230
Neisseria montereyensis CSL10203−ORH2
Neisseria zalophi ATCC BAA−2455
Neisseria dentiae DSM 19151
Neisseria musculi NW831
Neisseria weaveri NCTC12742
Neisseria arctica DSM 103136
Uruburuella suis DSMZ 17474
Uruburuella testudinis DSMZ 26510
Neisseria canis NCTC10296
Neisseria sp 83E034
Neisseria wadsworthii DSM 22245
Neisseria bacilliformis DSM 23338
Neisseria potus 3 SID 1128
Neisseria sp KEM232
Neisseria elongata ATCC 29315
Neisseria leonii 3986
Conchiformibius kuhniae DSM 17694
Conchiformibius steedae DSM 2580
Kingella denitrificans FDAARGOS 1060
Kingella bonacorsii Marseille−Q4569
Kingella oralis DSM 18271
Kingella kingae DSM 7536
Kingella negevensis SW7208426
Wielerella bovis CCUG 44465
Alysiella filiformis DSM 16848
Alysiella crassa DSM 2578
Simonsiella muelleri ATCC 29453
Eikenella corrodens NCTC10596
Eikenella exigua PXX
Eikenella halliae NML130454
Eikenella longinqua NML02−A−017
Neisseria shayeganii DSM 22244
Snodgrassella alvi wkB2 DSM 104735
Stenoxybacter acetivorans DSM 19021
Vitreoscilla massiliensis DSM 100958
Vitreoscilla stercoraria DSM 513

Supplementary Figure 9. Presence/absence, abundance and breadth of genomic coverage of *Neisseriaceae* species in metagenomic samples from two sites in the oral cavities of humans based on the sylph method. Eight samples from the Human Microbiome Project representing healthy individuals across two sites: the dorsum of the tongue and supragingival plaque. NCBI sequence read archive (SRA) sample accession numbers are provided in the figure for each metagenome. All *Neisseriaceae* species included in Chenal et al. are shown along with their phylogenetic relationships estimated by core genome analysis[24]. Circles are shaded according to the breadth of genomic coverage based on the read alignments and are scaled in size by the abundance in reads per kilobase-million (RPKM), based on the relative abundance calculated by slyph (-u option) and scaled by the total read count. Circles colored in purple are identified as present by the sylph tool.

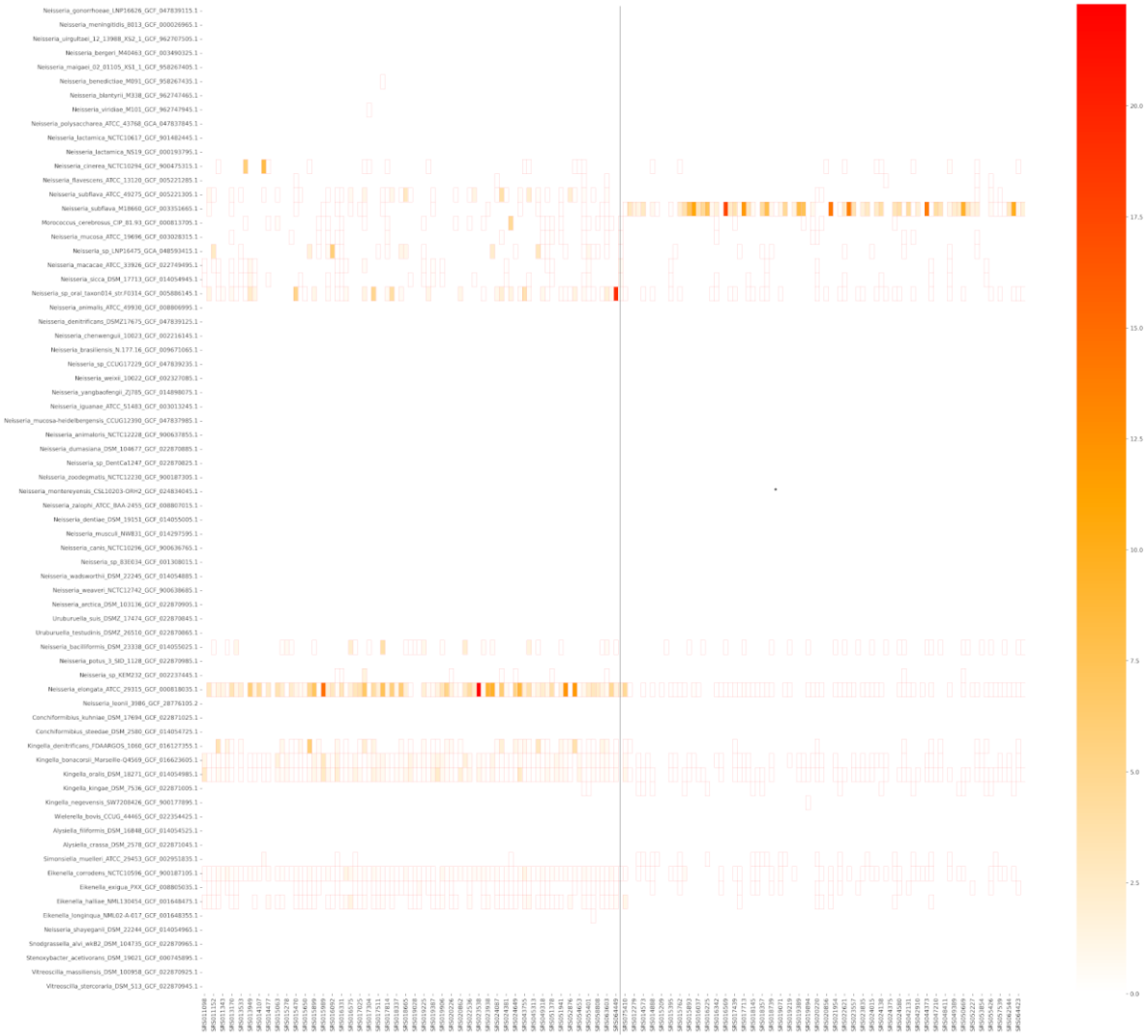


Figure S10. Presence/absence, abundance of *Neisseriaceae* species in metagenomic samples from two sites in the oral cavities of humans using the metapresence method with MAPQ20 alignment filtering. All samples from the Human Microbiome Project Phase I representing healthy individuals across two sites: supragingival plaque (left) and dorsum of the tongue (right), separated by a vertical line. NCBI sequence read archive (SRA) sample accession numbers are provided in the figure for each metagenome and in Table S5. Genomes present in a given sample are outlined with a box. Abundance is illustrated with the heatmap reads per kilobase-million (RPKM), based on the read count reported by metapresence/MAPQ20.

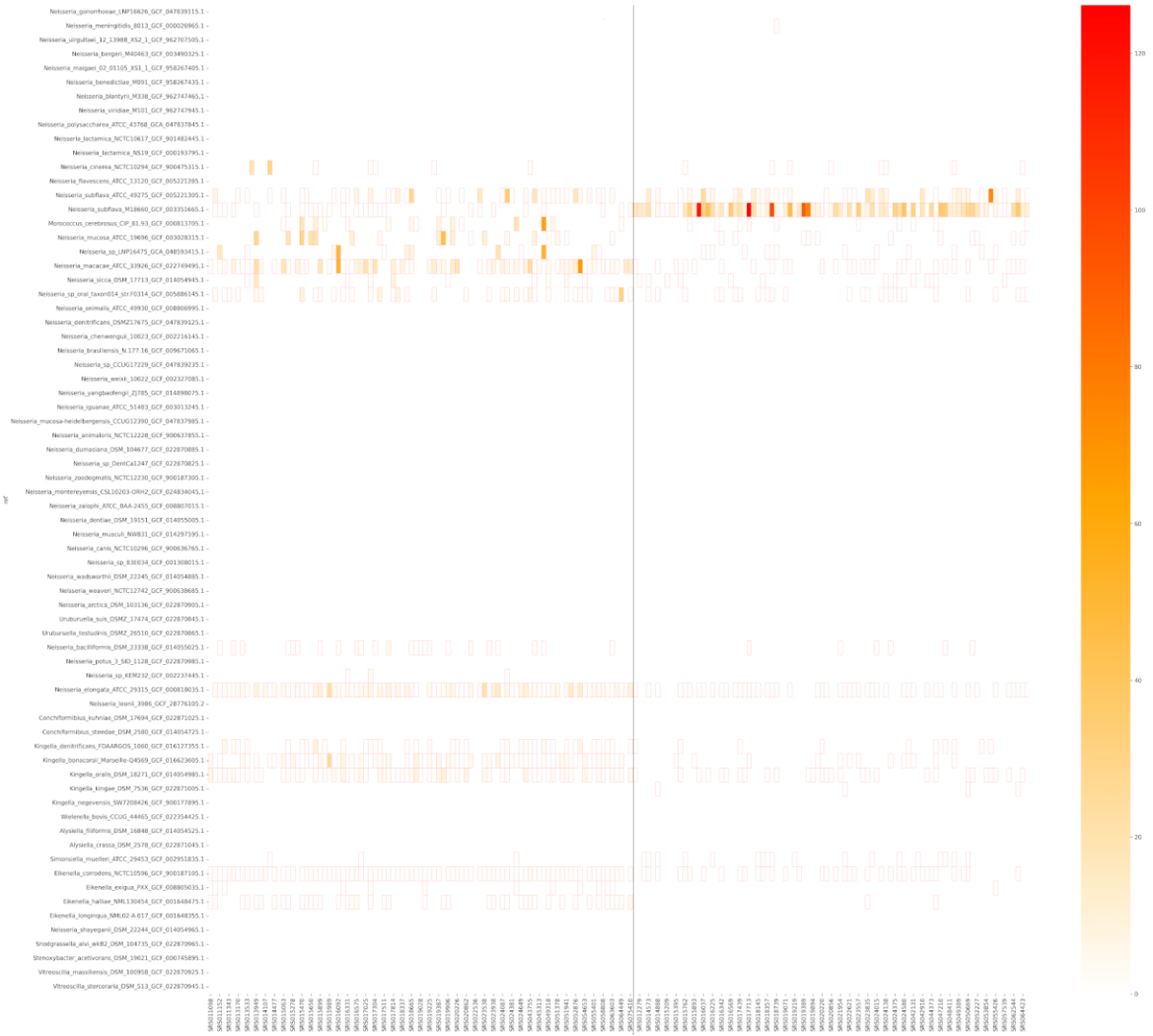


Figure S11. Presence/absence, abundance of *Neisseriaceae* species in metagenomic samples from two sites in the oral cavities of humans using the sylph method. All samples from the Human Microbiome Project Phase I representing healthy individuals across two sites: supragingival plaque (left) and dorsum of the tongue (right), separated by a vertical line. NCBI sequence read archive (SRA) sample accession numbers are provided in the figure for each metagenome and in Table S5. Genomes present in a given sample are outlined with a box. Abundance is illustrated by the heatmap in reads per kilobase-million (RPKM), based on the relative abundance calculated by slyph (-u option) and scaled by the total read count