Published online 10 May 2022

Nucleic Acids Research, 2022, Vol. 50, No. 13 ***e76***
https://doi.org/10.1093/nar/gkac294

How clear is our current view on microbial dark matter? (Re-)assessing public MAG & SAG datasets with MDMcleaner

John Vollmers*, Sandra Wiegand, Florian Lenk and Anne-Kristin Kaster*

Institute for Biological Interfaces 5 (Institut für Biologische Grenzflächen IBG 5), Karlsruhe Institute of Technology (KIT) 76344, Eggenstein-Leopoldshafen, Germany

Received January 31, 2022; Revised April 11, 2022; Editorial Decision April 12, 2022; Accepted April 13, 2022

ABSTRACT

As of today, the majority of environmental microorganisms remain uncultured and is therefore referred to as 'microbial dark matter' (MDM). Hence, genomic insights into these organisms are limited to cultivation-independent approaches such as single-cell- and metagenomics. However, without access to cultured representatives for verifying correct taxon-assignments, MDM genomes may cause potentially misleading conclusions based on misclassified or contaminant contigs, thereby obfuscating our view on the uncultured microbial majority. Moreover, gradual database contaminations by past genome submissions can cause error propagations which affect present as well as future comparative genome analyses. Consequently, strict contamination detection and filtering need to be applied, especially in the case of uncultured MDM genomes. Current genome reporting standards, however, emphasize completeness over purity and the *de facto* gold standard genome assessment tool, checkM, discriminates against uncultured taxa and fragmented genomes. To tackle these issues, we present a novel contig classification, screening, and filtering workflow and corresponding open-source python implementation called MDMcleaner, which was tested and compared to other tools on mock and real datasets. MDMcleaner revealed substantial contaminations overlooked by current screening approaches and sensitively detects misattributed contigs in both novel genomes and the underlying reference databases, thereby greatly improving our view on 'microbial dark matter'.

INTRODUCTION

Genomic information obtained through cultivation-independent sequencing techniques still remains the primary source of insight into the earths' uncultivated microbiome, the so called 'microbial dark matter' (MDM) (1–6). The continuous advancements in sequencing as well as (meta-)genome analysis methods have made this strategy nowadays widely accessible for a broad scientific community (7). This has led to an exponentially increasing amount of genome datasets of uncultured organisms in the form of 'metagenome-assembled genomes' (MAGs) as well as 'single-amplified genomes' (SAGs), both having different intrinsic advantages and disadvantages (4,8–10). MAGs are the result of so-called 'binning' approaches, which attempt to sort contigs (contiguously assembled sequence fragments) obtained from the combined genomic material of a diverse community into separate 'bins'. While each bin optimally represents the genome of an individual species (11), in reality a MAG is most likely a consensus genome gathered from all possible strain variants present in the sample (12) and generally excludes genomic islands and mobile genetic elements such as plasmids (13,14). Another prominent problem of this approach is the risk of assigning contigs from different species to the same bin, thereby constructing contaminated or chimeric MAGs (11,15). Single-cell genomics (SCG) can circumvent these problems (4). SAGs are the result of amplifying and sequencing DNA from individual cells which were physically separated from their community (2,3,8). In theory, such genomes are more reliable than MAGs, as they are directly derived from only a single organism. However, the biased nature of current whole genome amplification methods based on multiple displacement amplification (MDA) generally results in more incomplete SAGs than MAGs (11,16). In addition, SCG is highly susceptible to contamination such as traces of residual DNA left over from reagent production (17) or free environmental DNA fragments that may have been incidentally sorted together with the actual

*To whom correspondence should be addressed. Tel: +49 721 608 24236; Fax: +49 721 608 25546; Email: john.vollmers@kit.edu
Correspondence may also be addressed to. Anne-Kristin Kaster. Tel: +49 721 608 23005; Fax: +49 721 608 25546; Email: anne-kristin.kaster@kit.edu

© The Author(s) 2022. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/nar/article/50/13/e76/6583244 by KIT Library user on 23 December 2022

cell of interest (8). Furthermore, microorganisms that form tight aggregates or biofilms, may accidentally be co-sorted. As a result, both types of reconstructed MDM genomes share a common problem: the question of quality control. Since genomes obtained *via* both methods are typically highly fragmented (7,18), it is very hard to distinguish with confidence between correctly and incorrectly assigned sequence fragments without a pure reference culture.

An early solution to the problem was the estimation of completeness and contamination based on universal marker genes using the tools checkM (19) or Busco (20). Ever since the publication of recommended genome reporting standards represented by the 'minimum information about a metagenome-assembled genome' (MIMAG) and 'single amplified genome' (MISAG) (21), checkM has become the *de facto* gold standard for determining genome quality. While its widespread use has largely improved the quality of submitted genomes, this tool also has some operating principles with serious consequences: In a highly fragmented MAG or SAG, many contigs may not contain any conserved marker genes, making it impossible to reliably distinguish contaminants from 'correctly' assigned genome fragments. Offering different marker sets for different taxonomic levels, as implemented in checkM, might ease this classification issue a bit. However, the effectiveness of this approach is extremely limited in the case of under-sampled or uncultured taxa, for which little to no reference genome data is available. Furthermore, it must be kept in mind that checkM does not actually detect contamination directly, but rather uses a proxy metric to estimate it indirectly, that is, the multiplicity of assumed single-copy marker genes. Marker genes occurring more than once are interpreted as indicators for possible contamination without checking the phylogeny of said markers. On the one hand, this can lead to over-estimations since fragmented genes, paralogs, or closely related homologues might be classified as contamination (22), while on the other hand, actual contaminations may easily be missed, even if they encode conserved marker-genes, as long as this gene occurs only once in the analysed genome. Therefore, 'contamination' is a potentially misleading term for this metric, a more accurate term would be 'marker gene multiplicity'.

Other measures to identify and remove contaminating contigs from MAGs as well as SAGs have been undertaken in the past, but have not been universally adopted by the scientific community: Rinke et al. (3) analysed GC content distribution and kmer frequencies, as well as best blast hit classifications of the total encoded proteins for each contig. A similar approach is implemented in the tool refineM which was used to quality check 8000 MAGs obtained by Parks *et al.* from numerous metagenomes (9) as well as the tool ProDeGe which is provided by the Joint Genome Institute (23), both of which are, however, no longer maintained and supported. More recent and actively supported tools are MAGpurify and Gunc (24,25). MAGpurify was originally developed for the analysis of the human gut microbiome and consists of multiple modular approaches, such as the analysis of universal marker genes, GC content as well as kmer frequency profiles and the consideration of predetermined sets of known contaminants or trusted contigs (25). The MAGpurify reference database for classifying marker genes is based on MetaPhlAn2 (26), which excludes several non-cultured taxa. Gunc, on the other hand, currently does not support active decontamination of genomes but is designed to provide a robust and sensitive genome assessment in order to improve the current gold standard of checkM estimations (24). The underlying database for classifying protein coding genes is based on the Genome Taxonomy Database (GTDB) and therefore includes genomes of most currently known uncultured candidate taxa. Nevertheless, the high risk of introducing falsely classified sequences into reference databases and subsequent error propagation when submitting genome reconstructions of uncultured organisms demands a larger variation of independent contamination screening and filtering approaches. Otherwise, the potential for systematically overlooking of preventable contaminations due to unnoticed shortcomings and pitfalls of individual screening approaches becomes too high. Furthermore, effective classification, screening and filtering approaches also need to consider the ongoing problem of contaminations in public reference datasets (27,28) (see also Supplementary Information S1 and Supplementary Tables S1–S9).

Here, we present a new workflow as an alternative strategy for detecting and removing contaminations that is aware of potential reference database contamination, thereby minimizing the danger of error propagation. This workflow shows high sensitivity for contaminants even in highly fragmented genomes and in taxa that are underrepresented in public reference databases, making it equally applicable for prokaryotic MAGs as well as SAGs. We provide a free and open access python implementation of this workflow, called 'MDMcleaner', a contig classification and refinement tool. We also re-assessed the quality of presumed 'low contamination' MDM genomes in public datasets to elucidate how much our current view on the uncultured majority of microorganisms may be distorted by misattributed contigs in publicly deposited MAGs and SAGs. Furthermore, we illustrate potential problems in current best practice standards for genome assessments and propose a refinement of the current MIMAGs/MISAGs standards to reflect these problems.

MATERIALS AND METHODS

Python implementation of MDMcleaner

The MDMcleaner pipeline is implemented in Python 3, which allows its use as a standalone tool, a modular pipeline or as a python module. It requires Python 3.6+ and Biopython. All components were scripted with the general Unix principles in mind for easy piping. A general overview of the workflow is given below as well as in Figure 1. For a full description of the exact implementation for each step and the reference databases, please refer to Supplementary Methods S2 and Supplementary Tables S10–S12.

Basic workflow

Each genome is to be supplied as a fasta file of assembled contigs. Multiple hierarchically ranked levels of marker genes are extracted from the contigs, that is, small subunit

Downloaded from https://academic.oup.com/nar/article/50/13/e76/6583244 by KIT Library user on 23 December 2022

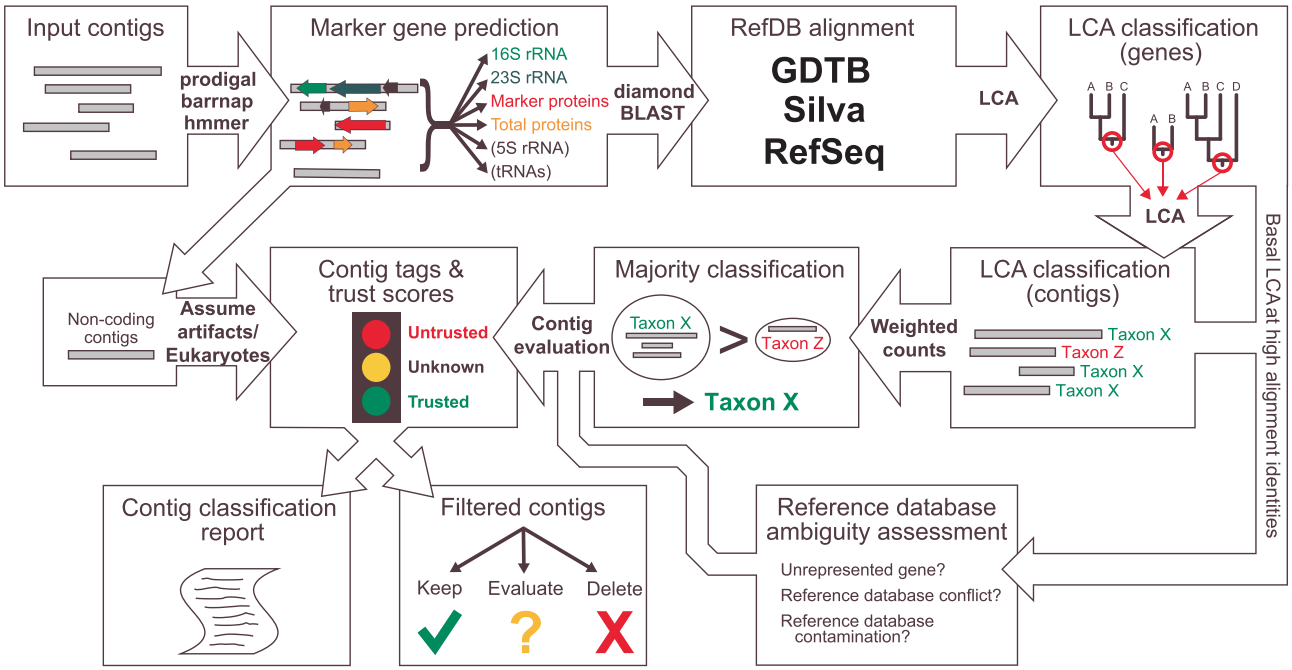

Figure 1. Basic MDMcleaner workflow. Ribosomal RNA and conserved as well as total protein coding gene sequences are detected on the input contigs and either aligned against a combination of GTDB, Silva and RefSeq references, or, in the case of non-coding contigs, indiscriminately assumed to be artefacts or eukaryotic contaminations. Alignments are then used to derive the least common ancestor (LCA) classification first for each gene, then for each contig and consequently the majority consensus annotation for the complete input genome. Potential reference database ambiguities are identified at these steps. Each contig is then evaluated and assigned a 'trustworthiness' score. In addition to a detailed contig classification report, separate fasta files are produced, evaluating the contigs to be kept, deleted or re-evaluated.

(SSU) rRNA genes, large subunit (LSU) rRNA genes, universal bacterial/archaeal protein coding marker genes, total coding sequences (CDS), and tRNA-genes (listed in decreasing hierarchy). Due to the general high coding density of bacterial genomes (29–31), non-coding contigs are considered artefacts or eukaryotic contaminants and therefore discarded.

All genes are then aligned against a reference database. In our implementation, an up-to-date database is derived from GTDB, SILVA and RefSeq (Supplementary Methods S1). This results in a condensed set of a few curated representative genomes per species, thereby avoiding bias from overrepresented, easy to culture taxa. For each gene, the respective blast-hits are filtered based on relative alignment score differences and used for preliminary taxonomic classifications using a least common ancestor (LCA) approach. Subsequently, contigs are then taxonomically classified via LCAs derived from the individual gene classifications of each marker-level, while subsequently keeping track of the respective average alignment identities. In order to avoid over-classification, each resulting taxonomic path is then pruned to ranks that are actually supported by the respective alignment identities, based on established and commonly used cut-off values for 16S rRNA (32,33), as well as protein coding genes (34,35) (Supplementary Methods S1).

At this point, LCA classifications may become apparent, that are limited to high taxonomic ranks (e.g. domain or phylum) despite consistently high alignment identities. Such cases can represent ambiguities or even contaminations in the reference database and may therefore require further downstream evaluation. The MDMcleaner pipeline includes a separate workflow for identifying likely reference database contaminations from such cases, which are then recorded in a blacklist. When provided to future runs, entries of this blacklist will be ignored during sequence comparisons, thereby preventing error propagation.

An overall genome classification is then derived as the weighted majority consensus of the corresponding individual contig classifications. Each contig is then assigned a 'trustworthiness' score based on how much its contig classification deviates from the overall genome classification and on which marker genes and corresponding alignment identities (if any) were involved. These scores range from 0 (lowest trustworthiness) to 10 (highest trustworthiness). Outputs of this workflow are detailed reports on each contig including the corresponding 'trustworthiness' score and a division of the contigs into separate fasta files containing entries to either keep, delete, or possibly evaluate further, respectively

MAG dataset selection

Due to the sheer number of MAGs available from the National Center for Biotechnology Information (NCBI) (currently containing > 100 000 genomes) a preselection was necessary. The NCBI assembly database was queried for entries of bacteria and archaea marked as 'metagenome derived' or 'derived from environmental sample' and not as 'derived from single cell amplification', and not marked as 'contaminated' or 'misassembled'. Focus was then narrowed onto MDM genomes by further limiting the selection to taxa which showed higher representation by MAGs

Downloaded from https://academic.oup.com/nar/article/50/13/e76/6583244 by KIT Library user on 23 December 2022

and SAGs than by NCBI RefSeq entries and which contained <500 RefSeq entries in total, resulting in a preliminary dataset of >50 000 genomes. In order to focus on the currently most trusted MAGs, this dataset was then reduced to 'high quality' MAGs based on current MIMAGs standards (checkM completeness estimation ≥ 90% and a marker gene multiplicity ≤ 5%). This resulted in a final analysis set of 4011 presumed high quality prokaryotic MAGs from predominantly uncultured taxa. The corresponding NCBI accession numbers are provided in Supplementary Table S1.

SAG dataset selection

Of the 1667 prokaryotic SAGs publicly available from NCBI (as of November 2020), 1597 displayed <5% marker gene multiplicity during checkM analyses, indicating low contamination and qualifying for detailed screening with MDMcleaner. Of this selection, 149 were >90% complete, 677 were >50% complete and 772 showed <50% completeness qualifying as 'high quality', 'moderate quality' and 'low quality' genomes by current MISAG standards, respectively (21). However, since MDA bias may make it currently impossible to capture full genomes for some taxa, especially in the case of high GC organisms (36), it is possible that some taxa would be better represented by an uncontaminated but incomplete SAG than a partially contaminated but seemingly complete MAG. Therefore, MISAG quality terms were of less significance for the selection of the datasets. Instead, SAGs of all completeness values were analysed, as long as the contamination estimates represented by checkM marker gene multiplicity were below the 'high quality genome' cut-off of 5%. The corresponding NCBI accession numbers are provided in Supplementary Table S2.

Benchmarking datasets

257 genomes of isolates representing novel taxa at least on genus, and even up to phylum level were downloaded from NCBI RefSeq to ensure that these were not already represented in the applied reference database and to most accurately represent 'microbial dark matter' (Supplementary Table S3). The genomes were then randomly cut into fragments representing contig size ranges realistic for SAGs and MAGs (0.2–20 kb, with a median length of 10 kb). 25% of each genome was then randomly replaced with fragments of all other 256 genomes as well as fractions of the human genome, in order to produce mock bins with a known completeness of ∼75% and known contamination fractions of 25%, each. Size-skewed mock genomes were generated from a smaller subset of four of the above mentioned isolate genomes and were also cut to fragment sizes between 200 bp and 20 kb, but were skewed towards a specific major fragment length by addition of contaminating fractions of 25% obtained from the other mock genomes cut to specific size ranges. This resulted in 180 size-skewed mock genomes (Supplementary Table S4). Statistical testing to compare the contamination averages gained from the used tools was either done with Welch's *t*-test or Welch's ANOVA with post hoc Games-Howell pairwise comparison. Welch's testing was selected because all variances were previously determined to be unequal by a Levene's test. *T*-tests were used for the pairwise comparison of two data sets and ANOVA testing was done for the comparison of more than two data sets. In order to assess performance on eukaryotic contaminants the same four bacterial genomes were chosen, however, this time human genome fragments were used as contamination (Supplementary Table S5).

RESULTS AND DISCUSSION

The initial situation

The genome analysis of uncultured bacteria remains a constantly evolving field due to frequent and drastic improvements in sequencing and bioinformatics methods. Nonetheless, for the 'darkest' parts of microbial dark matter (MDM), especially the so-called 'candidate phyla' with not a single cultured representative available, MAGs and SAGs are currently our only source of genomic information (3,6,8,37,38). Therefore, high caution is advisable when deriving general conclusions from such datasets, as any residual contaminations in these genomes could greatly bias our assumptions on their metabolic properties and evolution. The MDMcleaner pipeline has been developed based on experiences gained during past MAG and SAG analyses (2,8,22,39), and is designed specifically to combat this issue: contaminations in MAG and SAG datasets as well as reference databases (Figure 1). This workflow maximizes the reference genome information used for MAG and SAG evaluation by integrating the GTDB and SILVA databases as well as curated eukaryotic and viral reference datasets from NCBI RefSeq. It also maximizes the analysable sequence information by applying multiple marker gene levels in decreasing priority (rRNA, conserved marker proteins and total proteins, respectively). Furthermore, the MDMcleaner workflow utilizes a least common ancestor (LCA) contig classification approach that is aware of reference database contamination and is therefore resistant to error propagation by contaminated reference genomes (see Material and Methods).

The current genome reporting standards for MAGs and SAGs, designated MIMAGs and MISAGs, respectively, define up to five percent contamination as acceptable for 'high quality' genomes (21). The exact fraction of contamination reported for a given genome and therefore the question which genomes exactly exceed this cut-off is highly subjective, as it is dependent on the method of contamination detection (Figure 2A). Today, the tool checkM represents the *de facto* gold standard for such anaysis (19,21). MAGpurify excludes a smaller number of MAGs from the MIMAGs 'high quality' status than checkM. Gunc, on the other hand, identifies much larger fractions of contaminating contigs in the analysed genomes, with the exact fraction depending on the considered taxonomic levels, effectively including fewer genomes within the 5% contamination cut-off. All three tools appear to agree that SAGs generally display smaller fractions of contaminating contigs than MAGs, which also explains the rising popularity of single-cell genomics for analysis of MDM. Counter-intuitively, the analysed prokaryotic domain also appears to have a strong influence: checkM tends to report a smaller portion

Downloaded from https://academic.oup.com/nar/article/50/13/e76/6583244 by KIT Library user on 23 December 2022

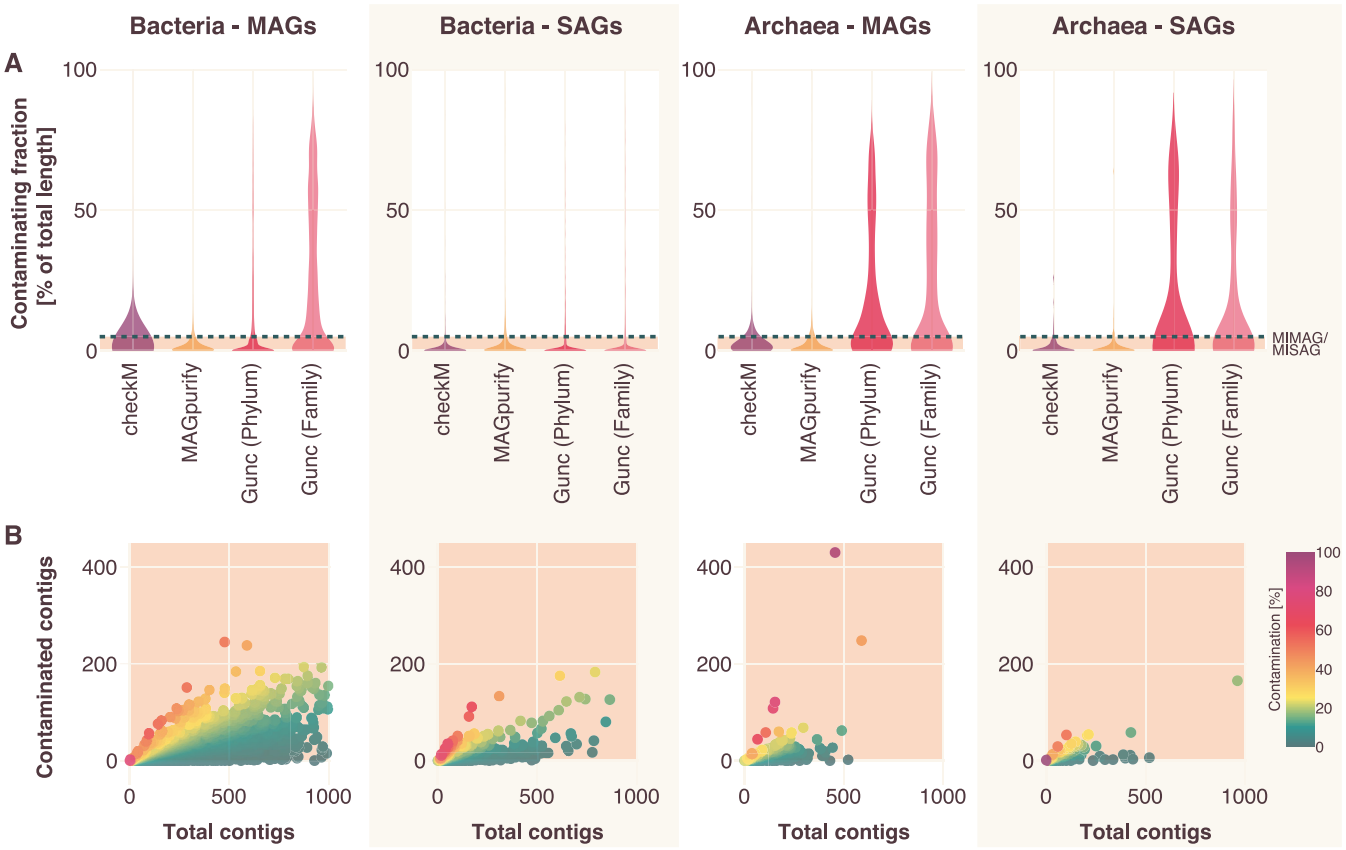

Figure 2. Current assessments of database contaminations. (**A**) Distribution of contamination estimations by different tools for bacterial and archaeal MAGs and SAGs, respectively. The current MIMAG/MISAG 'high quality' cut-off is marked by a horizontal dotted line. The fraction of genomes fulfilling 'high quality' criteria is highly dependent on the applied assessment tool and corresponding settings, but also on the genome type and microbial domain. (**B**) MDMcleaner re-assessment of assumed 'high quality' genomes, displayed as scatterplots, plotting the number of contaminated contigs against the number of total contigs per analysed genome. Contamination fractions determined by MDMcleaner are additionally indicated by colour, as per colour code on the right side. Significant fractions of potentially contaminating contigs are found, even in genomes that are considered 'high quality' based on checkM assessments and current MIMAGs/MISAGs standards

of archaeal MAGs as contaminated than bacterial ones, whereas the exact opposite is true for Gunc. This crucial detail is indicative of another subjective factor influencing genome assessments: the representation of the applied reference databases. Since there are far fewer cultured representatives available for archaea than for bacteria, the reference set for classifying archaeal genomes is generally smaller. Gunc makes use of the GTDB reference database (24,40), which includes curated genomes from uncultured taxa. CheckM categorically excludes MAG and SAG datasets as references (19), thereby specifically discriminating against under-cultured taxa. MAGpurify marker gene classifications are based on the MetaPhlAn2 database (25,41), which was originally compiled in 2012 and is much smaller than the representative genome database of GTDB, with 17 000 genomes compared to more than 32 000 (and even >250 000 total) GTDB genome entries. Consequently, it seems likely that checkM and MAGpurify may mis- or under-classify archaea, leading to a potential underestimation of the contamination fractions for genomes in this domain. The same problem is most likely to also affect bacterial MDM, such as underrepresented bacterial candidate phyla. When genomes that were incorrectly assumed to be 'high quality' on this basis are then added to reference databases, future genome assessments are further affected due to error propagation caused by misclassified reference sequences. The ultimate consequence is a highly problematic dilemma for MDM genomics: The fewer cultured representatives there are available for a taxon, the more reliant the scientific community becomes on MAG and SAG datasets, while at the same time these datasets become less and less reliable.

We therefore decided to specifically re-examine the presumed 'high quality' and/or 'low contamination' MAG and SAG datasets with our revised MDMcleaner workflow. Interestingly, MDMcleaner screenings of these subject genomes indicated frequent instances of far >5% contamination. Occasionally, even more than half of the contigs in the assembly showed problematic assignments, likely representing contaminations (Figure 2B). This short analysis questions the trustworthiness of current MDM evaluations and clearly illustrates the need for stricter and more sensitive contamination filtering procedures. Furthermore, the question remains whether up to 5% contamination can be truly assumed to be likely 'false positives' and therefore tolerable for submitted reference genomes. To answer this, we

Downloaded from https://academic.oup.com/nar/article/50/13/e76/6583244 by KIT Library user on 23 December 2022

validated the MDMcleaner workflow and other contamination assessment tools on known reference datasets before further examining our findings on contaminated MAGs and SAGs, as presented in the following sections.

Validating sensitivity and precision of the MDMcleaner workflow

MDMcleaner was validated on genomes obtained from isolates, which are more likely to be free from contamination. For this purpose, 257 isolate genomes not yet present in the applied version of the GTDB database were selected, all of which are novel at least on genus level in order to most closely mimic MDM. This also includes instances where multiple species represent the same novel genus in order to reflect all levels of inter-taxon homologies. The genomes were randomly cut into fragments representing contig size ranges realistic for SAGs and MAGs (0.2 to 20 kb, with a median length of 10 kb) and 25% of each genome was randomly replaced with fragments of the other 256 genomes as well as fractions of the human genome, in order to produce mock bins with a known completeness of ∼75% and known contamination fractions of 25%, respectively.

A preliminary checkM analysis showed that the completeness was predominantly correctly estimated with a median value ∼77%. The checkM 'contamination' metric based on marker gene multiplicities, however, largely underestimated the actual contamination fraction with a median value of ∼14% and a large outlier range of 0–93% (Figure 3A). The average contamination fraction reported by MDMcleaner, however, was close to the actual value of 25%. In contrast, MAGpurify routinely underestimated the contaminating fraction with a median value of 9%. Gunc correctly assessed all mock genomes as contaminated, based on unmistakably high 'Cluster separation score' (CSS) (24) values averaging at a median of 0.98 (Supplementary Table S3). The actual proportion of contamination in each genome was, however, misjudged by Gunc with estimates ranging from 3 to 61% (median: 11%). This indicates that Gunc may be a good indicator for deciding whether a genome is likely to be contaminated, but not well suited for filtering the affected contigs. This observation is also represented in the fact that Gunc, similar to checkM, is currently being provided as a contamination assessment, not as a filtering tool. ANOVA testing proved that the observed differences were not random, with P values below 0.0001 ('extremely significant').

For MDMcleaner and MAGpurify, results were further analysed regarding true and false positives, reflecting sensitivity and precision, respectively. MDMcleaner demonstrated exemplary sensitivity with median true positive rates of 90% (Figure 3B). In comparison, MAGpurify identified significantly fewer true positives with a median of only 34%, illustrating that a large fraction of potential contamination would be missed by this screening tool. Regarding precision, MDMcleaner appeared to display a larger range of false positive fractions than MAGpurify on these test sets: MAGpurify yielded only up to 3% false positives, with a median value of 0.2%, while the respective fractions returned by MDMcleaner averaged at a median value of 1.6%, with some isolated extreme cases even reaching up to 27%. Both observations were tested via Welch's t-test, with P values <0.0001 ('extremely significant'). Most false positives in these cases represented taxonomic conflicts below class or order level (Supplementary Table S3), showing that taxa on domain and phylum levels were robustly assigned. The few extreme cases in which a significant number of phylum-level conflicts were found, mostly represented taxa which were uniquely reorganized on phylum level within the GTDB taxonomy (e.g. the splitting of the former Firmicutes into five separate phyla designated 'Firmicutes' and 'Firmicutes_A-D') which are not yet officially recognized by the broader taxonomic community, possibly indicating the need for future fine-adjustments of the GTDB taxonomy. Alternatively, it may also be possible that these taxa possess a higher genome plasticity and therefore more inter-phyla homologies than most other phyla.

Since actual metagenome and single-cell genome assemblies are often skewed towards short contig lengths (18), a separate analysis was performed in order to elucidate the exact influence of contig size on the efficiency of contamination detection. For this, a smaller genome subset was selected consisting of *Atribacter laminatus* RT761 (42) (as the first and only isolate of its phylum and therefore a good representative for actual MDM), and three additional randomly selected reference genomes, each representing different bacterial phyla. As above, the genomes were again randomly cut to contig size ranges of 200 bp to 20 kb in triplicates. Contaminant contigs were cut into five different size ranges and added to produce a total set of 180 differently size-skewed prokaryotic mock genomes. Again, MDMcleaner correctly identified the vast majority (>90%) of true positive contaminants, displaying extraordinary sensitivity, while maintaining relatively low false positive rates (Figure 4, top panel). In contrast, MAGpurify appeared to be increasingly biased when contig size distributions were skewed towards shorter contig lengths between 200 bp and 2 kb, with the major source for misclassifications by MAGpurify being GC content analyses. A point where MAGpurify achieved significantly lower false positive rates than MDMcleaner was only reached when contig lengths were skewed towards large sizes of 5 kb and above (Figure 4). The true positive rate of MAGpurify on the size-skewed mock datasets on the other hand, remained consistently below those yielded by MDMcleaner even at large contig sizes. This is a significant advantage of MDMcleaner regarding both precision and sensitivity for analysing highly fragmented MAGs and SAGs.

When closely examining the taxonomic divergencies observed at different taxonomic levels, the majority of false positives reported by MDMcleaner occur at species and genus level. At such lower levels, taxon-delineation becomes less exact, partly due to the fact that horizontal gene transfer is more common between closely related organisms (43), but also due to the fact that different clades are not evenly represented in reference databases. This is also reflected in the fact that Gunc, by default, performs contamination assessment only down to genus level (24). The vast majority of true positives, however, were based on contig annotations on family level or below, which was therefore selected as the default cut off level for MDMcleaner. With this cut off, MDMcleaner yielded false positive rates that consistently

Downloaded from https://academic.oup.com/nar/article/50/13/e76/6583244 by KIT Library user on 23 December 2022

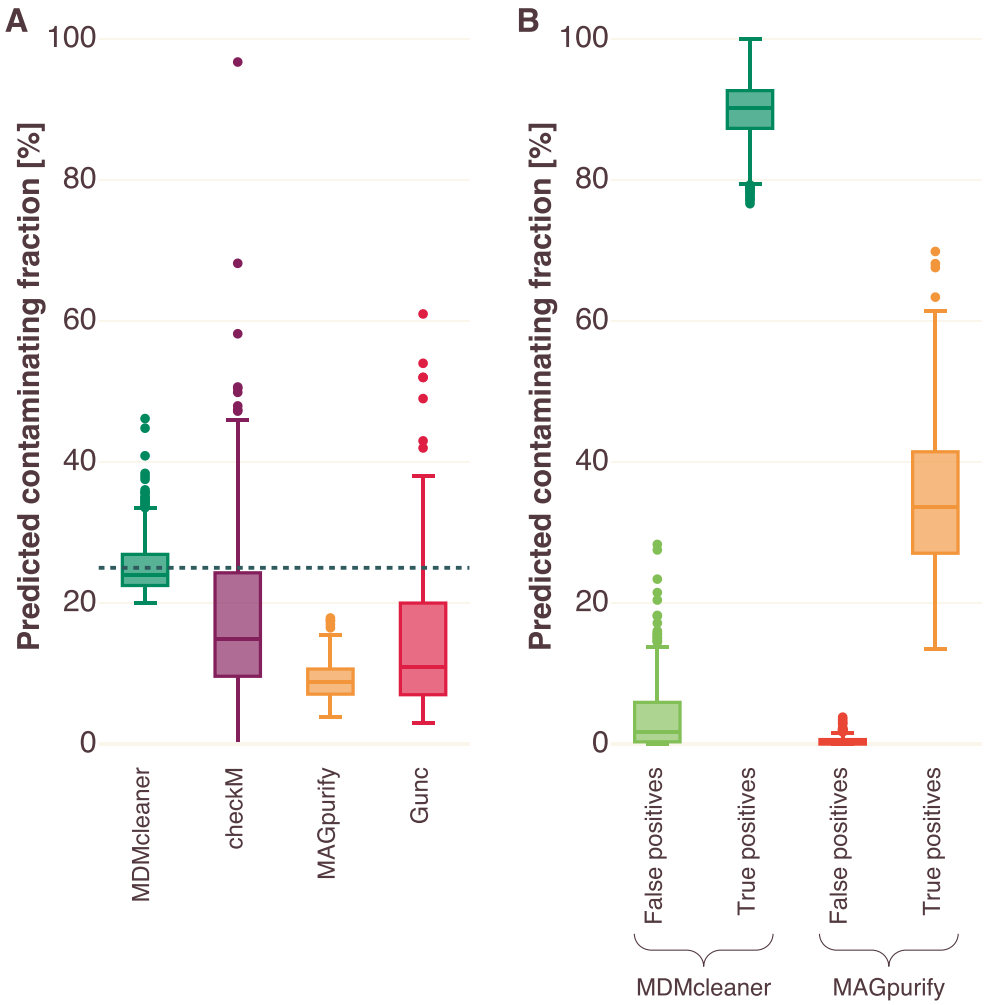

Figure 3. Benchmarking the MDMcleaner workflow on mock microbial dark matter SAGs. (**A**) Overall contamination fractions reported by different tools. The boxplots show the distribution of contamination values reported among the testgenomes by each tool, regardless of actual true positive or false positive rates. On average, MDMcleaner reported contamination values closest to the actual known contamination rate of 25%. CheckM, MAGpurify, and Gunc tended to underestimate the contamination fractions. (**B**) Detailed distribution of the false positive and false negative screening results among the test genomes, returned by MDMcleaner and MAGpurify. Since checkM and Gunc do not return the actual assumed contaminants, these tools were omitted from this analysis. MDMcleaner identified almost all contaminant contigs, with more than twice the success rate compared to MAGpurify.

averaged at a median value of 1.6%, even at very short contig lengths.

Because a massive influence of eukaryotic contaminations on curated prokaryotic reference datasets has been reported (44), a third and final benchmark was performed to gauge the effectiveness of our workflow and other tools to eliminate eukaryotic contaminants. For this purpose, the test genomes were purposely contaminated exclusively from human genomes. However, in reality eukaryotic contaminants in microbial metagenome analyses may stem from multiple unsuspected or at least unknown sources. For this reason, although it is possible to supply the human genome as a 'known contaminant' to MAGpurify for targeted screening purposes, this option was not used in order to assess how well each tool can identify unsuspected eukaryotic contaminations under standard settings. Under these conditions, some cases of human DNA contaminated mock genomes actually passed the Gunc assessment, indicating that Gunc may not be as reliable for detecting contaminations from eukaryotic sources as from prokaryotic sources. MDMcleaner also showed a 5× higher sensitivity than MAGpurify (Supplementary Figure S1). Furthermore, false negative contigs missed by MDMcleaner were almost exclusively 'unclassified' (thereby yielding a low trust value of five), indicating that CDS were predicted on this contig but did not yield significant BLAST results for meaningful LCA classifications. Such spurious prokaryotic CDS predictions are likely to occur from time to time on longer non-coding eukaryotic DNA stretches. In fact, such erroneous CDS assignments on contaminating eukaryotic contigs in bacterial genomes have already led to the assignment of entire spurious 'conserved protein families' in the past (44). However, in the case that eukaryotic contaminants are a likely problem, the resulting unclassified contigs can be specifically extracted from the MDMcleaner output and subjected to nucleotide level alignments against eukaryotic reference genome databases (this was omitted from the current python implementation due to the large

Downloaded from https://academic.oup.com/nar/article/50/13/e76/6583244 by KIT Library user on 23 December 2022

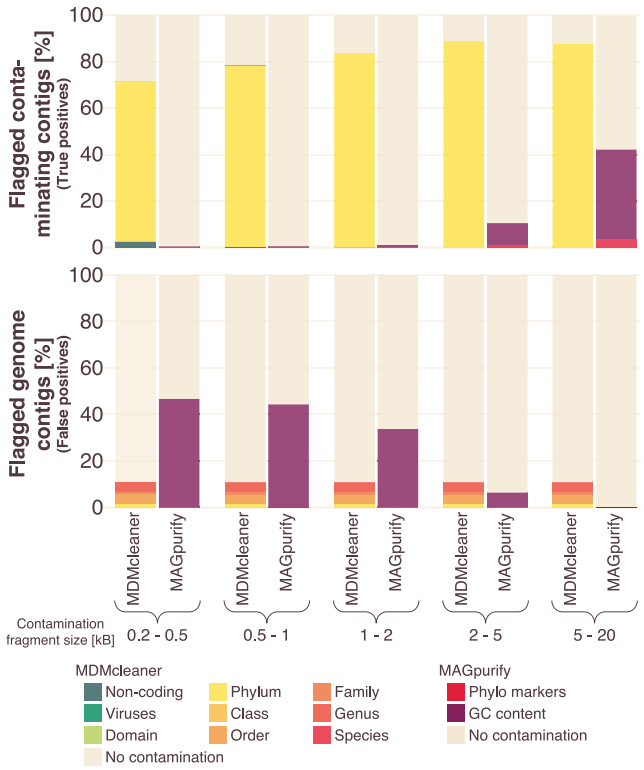

Figure 4. Influence of contig size and screening categories. Barcharts show average fractions of contaminant contigs and genome contigs classified as 'contamination' (representing 'true positives' and 'false positives', respectively) among the test genomes at different contig size distributions. The corresponding taxonomic level (MDMcleaner) or assessment metric (MAGpurify) that caused the respective contaminant classifications is indicated by the colour code below. MDMcleaner yielded consistent high true positive and low false positive rates, regardless of average contig size. Potential false positives were predominantly based on species and genus level classifications, which hardly contributed to the true positive fractions. MAGpurify showed optimal results only at contig sizes larger than 5 kb, with low true positive but high false positive rates, predominantly based on GC content analyses at lower contig sizes. No results were obtained for the MAGpurify screening category 'tetramer frequencies', which is therefore not shown here.

size of the additionally required reference databases), or assessed with dedicated eukaryotic metagenome classifiers such as Tiara, EukRep or Whokaryote (45–47). In summary, MDMcleaner outperformed other currently available tools in this benchmarking approach and provided exemplary sensitivity for contaminations, while having a comparatively low false positive rate on fragmented query genomes such as MAGs and SAGs.

Assessment of current microbial dark matter genomes: how clear is our view?

A general overview of the MDMcleaner findings in the analysed 6508 MAG and SAG datasets are presented in Figure 5. Potentially problematic contigs reported by MDMcleaner can be divided into one of the three following categories: (A) 'non-coding' contigs, (B) 'divergent taxonomy', which can range from virus/domain to species level (with divergence above family level being however ignored by default) and (C) 'reference database ambiguities' (representing potential contaminations within the reference database but not necessarily within the analysed MAG & SAG datasets) (Figure 5A). Non-coding contigs, which do not encode any detectably prokaryotic RNA gene or CDS are usually not considered by other screening tools. However, due to the comparatively high average coding density of prokaryotes (29–31), they may be seen as indicative for either eukaryotic contamination or amplification artefacts, such as primer-dimers (48). Such non-coding contigs occur drastically more often in the analysed SAGs than MAGs (Figure 5A), thereby most likely representing MDA artefacts. Taxonomic divergence, on the other hand, indicates a contig with a taxonomic classification that does not match the majority classification of the corresponding genome. The exact rank at which this divergence occurs is relevant for estimating the likelihood of the contig being an actual contaminant as regions of high identity are more likely to be shared between related species than separate phyla, as already indicated in the benchmarking tests (Figure 4). Accordingly, divergences on genus and species level were generally more frequently observed within the analysed MAGs and SAGs than on class to family level, and are, by default, ignored during contamination filtering by MDMcleaner. By far the most significant potential contaminations seem to occur at phylum level. In archaeal genomes, especially archaeal MAGs, even divergences on domain level are frequent. Since the knowledge on archaeal genomics is even more limited compared to their bacterial counterparts, such substantial contaminations could drastically distort our view on the entire archaeal domain.

Examining the exact phylogenetic composition of contaminants detected on high-ranking domain or phylum levels shows that the most common source of potential domain level contaminants in archaea are of presumed bacterial origin (Figure 5B). Of course, this may partially be simply caused by the lower representation of archaea in reference databases compared to bacteria, resulting in a possible bias towards bacterial classifications. However, since almost half of these could be further classified to lower ranks even down to genus and species and furthermore, since the fraction of bacterial contigs is far higher in the MAGs than at the SAGs level, it appears evident that archaeal dark matter genomes actually contain a substantial number of bacterial contaminants. Interestingly, except this general influence of bacterial contaminants in archaeal genomes, no clear trend of specific systematic contaminant sources is visible within the MAG and SAG datasets (Figure 5B). In fact, the potential contaminants appear to originate from a huge variety of different source organisms, with each affected genome possessing a more or less unique composition of potential contaminants. This is of high relevance for the interpretation of SAGs, as it stresses the need to apply strict contamination assessment and filtering approaches beyond simply screening for known MDA contaminants: With no clear dominance of specific individual contaminants, it appears unlikely that residual contaminants from the manufacturing process of MDA reagents (17) are the only contaminant source here. Likewise, it seems unlikely that unsterile work conditions are the major contamination source, as this would likely result in a dominance of contaminants originating from common lab organisms or human cells. Instead,

Downloaded from https://academic.oup.com/nar/article/50/13/e76/6583244 by KIT Library user on 23 December 2022

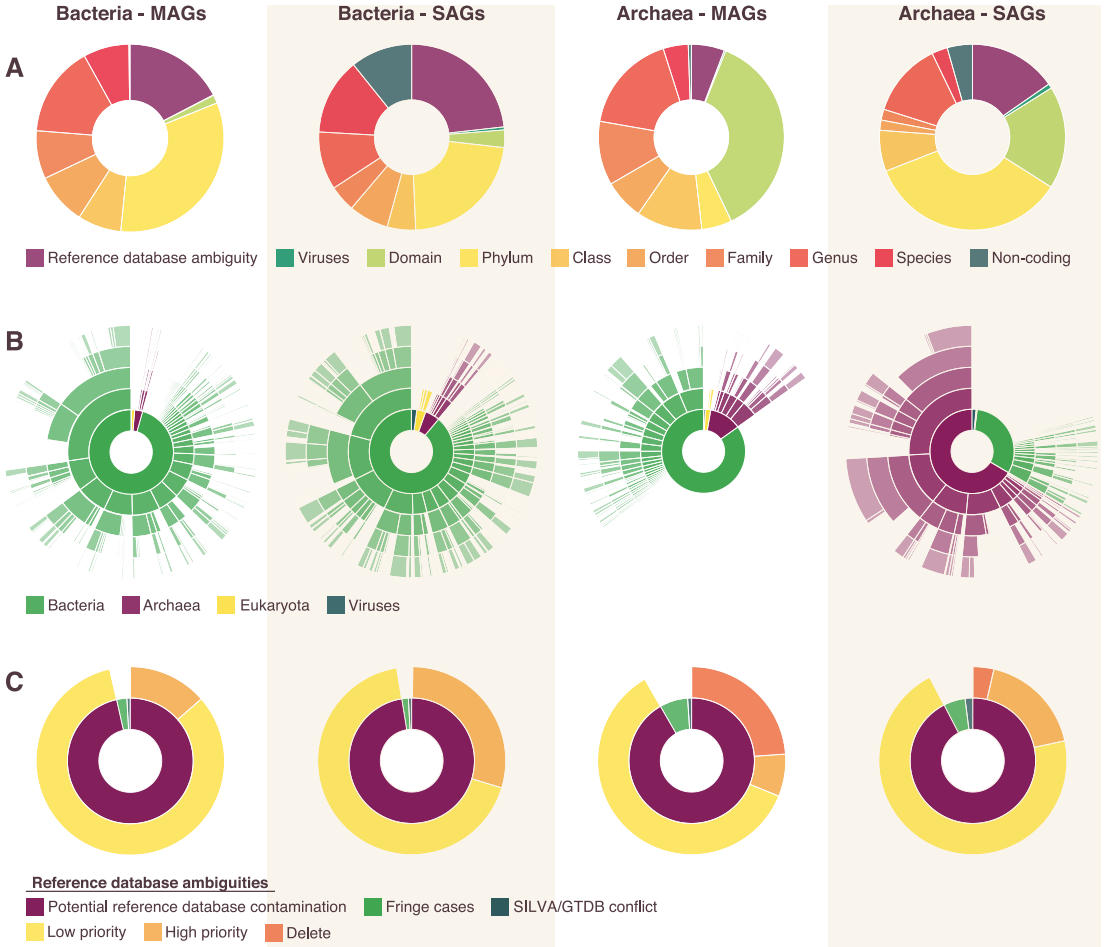

Figure 5. Detailed results of contig classifications by MDMcleaner on public MAGs and SAGs. (**A**) Categories of 'low trustworthy' and therefore potentially problematic contigs. Such contigs may represent reference database ambiguities, viral sequences, taxonomic divergences on domain to species level or non-coding contigs, as indicated by the colour code below. High fractions of reference database ambiguities, potentially representing reference database contaminations, are reported in all cases, while large fractions of non-coding contigs are far more prevalent in SAGs than in MAGs. (**B**) Radial charts showing the diversity of the most obviously contaminating contigs (i.e. with divergent taxonomic classification at domain or phylum level) in the analysed genomes. The innermost ring represents classifications on domain level, while the outermost represents classifications up to species level. Different domains are indicated by different colours. Apart from an apparent prevalence of bacterial contaminations in archaeal genomes, no clear overrepresentation of systematic contaminants can be recognized. (**C**) Breakdown of detected reference database ambiguities into the categories 'Potential reference database contaminations' (showing moderate to high identity blast hits to different phyla), 'Fringe cases' (showing hits to multiple phyla, but only at relatively low identities) or 'SILVA/GTDB conflicts' (representing differences in the GTDB and SILVA taxonomic systems). Potential reference database contaminations can be further distinguished into fractions requiring 'low priority' or 'high priority' evaluation or even directly categorized as 'delete' based on average blast identities to different reference genomes.

it appears that contaminants originate from the respective environmental sample that was sorted, e.g. caused by accidentally sorting multiple cells or by capturing free environmental DNA fragments together with individual cells (8).

Despite low checkM marker gene duplicity values, the MIMAGs/MISAGs high quality cut-off of up to 5% contamination (21) was exceeded in 15% of the MAGs and 12% of the SAGs according to MDMcleaner results (Supplementary Tables S1 and S2). In extreme cases, MDMcleaner estimations even reached 48% contamination for MAGs and 69% for SAGs. The most contaminated MAG from the MDMcleaner results is *Nitrosopumilaceae* archaeon Plut_88877 (NCBI acc. no. GCA_012271085.1) (49) (Table 1). This genome displayed very good checkM assessment metrics of 96-99% completeness and marker gene multiplicity of 3.41%, officially qualifying the genome as 'high quality' and understandingly giving the researchers little reason to perform additional decontamination steps. In this particular example, the majority of the presumed contaminant contigs detected by MDMcleaner were classified as bacterial and therefore appear to be misattributed on domain level (Supplementary Figure S2A), an assessment that is also strongly supported by BLAST alignments of encoded marker genes against the NCBI RefSeq databases. Furthermore, this finding is also strongly supported by both Gunc and MAGpurify (Table 1).

The most contaminated SAG, with >69% contamination is a presumed Rhizobiales strain (NCBI acc. no. GCA_000464375.1) which, however, actually appears to be a 'Frankenstein' mix of contigs from different organisms such as Proteobacteria, Bacteroidetes, Verrucomicrobia and Planctomycetes (Supplementary Figure S2B). This

Downloaded from https://academic.oup.com/nar/article/50/13/e76/6583244 by KIT Library user on 23 December 2022

Table 1. Most contaminated genomes of different types and completeness categories, according to MDMcleaner. All genomes fulfil the MIMAG/MISAG contamination criteria for 'high quality' genomes based on checkM, but GCA_000464375 would be nonetheless considered 'low quality' due to the lack of universal single copy marker genes causing low completeness

Genome info		MIMAG/	checkM		MDMcleaner		Mpurify	Gunc	Taxonomic classification		
Assembly acc	type	MISAG quality	compl.	contam	ambig	contam	contam	CSS	Domain	Phylum	Lowest taxon classification
GCA_012271085	MAG	high	96.26	3.41	0.1%	48.0%	53.0%	0.99	Archaea	Thermo-proteota	Nitrosopumilaceae (family)
GCA_000464375	SAG	low	0	0	5.7%	69.3%	17.4%	0.94	Bacteria	Alphaproteo-bacteria	Sphingomonas (genus)
GCA_000510525	SAG	high	97.44	1.72	0.2%	3.5%	3.4%	0.65	Bacteria	Bacteroidota	Tannerella (genus)

MAG even contains multiple rRNA genes of different organisms, including unambiguous planctomycetal and actinobacterial 23S rRNA gene sequences. Gunc and MAGpurify also yielded high contamination values for this genome, with MAGpurify reporting a drastically lower, albeit still notably high, contamination estimate of 17%. Since no universal single copy marker genes could be found, the checkM assessment of this genome yielded 0% completeness estimates and hence 0% marker gene multiplicity. This makes this genome appear pure, although obviously incomplete. In fact, observed discrepancies between low checkM marker gene duplicity values and high MDMcleaner contamination reports, generally increased with decreasing genome completeness estimates (Supplementary Table S2). This indicates that checkM tends to increasingly underestimate contaminations the less complete the analysed genomes are. Similar observations reported by Becraft et al. (16) support this assessment. However, this directly contradicts current MIMAGs/MISAGs reporting standards, which allow for even higher contamination values at lower genome completeness levels (21).

Curiously, the presumed Nocardiodes SAG is far from being the only case of publicly deposited genomes containing misattributed ribosomal RNA contigs. We identified 381 cases among the herein analysed SAGs and MAGs alone (Supplementary Table S6). Furthermore, reference database ambiguities detected during the corresponding MDMcleaner analyses indicate the presence of at least 175 instances even among representative GTDB reference genomes (Supplementary Table S7). The 56 most striking cases of almost full length 16S rRNA genes misattributed at phylum level or above are illustrated in Supplementary Figure S3.

When focusing exclusively on SAGs showing >90% completeness (qualifying as 'high quality genomes'), the highest observed contamination was only 3.5% (Table 1, Supplementary Figure S2C) for an uncultivated *Tannerella* member (NCBI acc. no. GCA_000510525). Incidentally, similar contamination values were also reported by MAGpurify and Gunc for this genome, showing that this fraction is unlikely to simply be the result of spurious false positive contaminant detection but instead represents actual contamination. While such a relatively small fraction may seem negligible in comparison to the higher contamination rates observed in numerous MAGs, it nonetheless represents avoidable contamination that will cause misleading comparison results and can lead to error propagation, thereby exponentially increasing its effect on gradual reference database corruption. Overall, MDMcleaner reported at least minimal contamination fractions in the majority (70%) of MAGs but only 44% of the SAGs (Supplementary Tables S1 and S2). This severe difference indicates that, despite the involved contamination sensitive MDA procedure, SAGs are generally less prone to misattributed contigs than MAGs. This is especially noteworthy when considering that, in contrast to the analysed MAGs, the majority of the herein analysed SAGs would not have been categorized as 'high quality' by current MISAG standards due to lower genome completeness estimations.

In direct comparisons of the respective analyses results, MDMcleaner does not yield larger outlier and interquartile ranges than MAGpurify or Gunc, showing that the presumed slightly higher false positive rate indicated in the benchmark tests does not lead to systematic overestimations of contaminations in actual genome data (Supplementary Figure S4A). In fact, much greater extreme values were observed by MAGpurify, even reaching an implausible 100% in the case of two MAGs (GCA_003712165.1 and GCA_014762685.1). Close inspection showed that these were almost exclusively classified based on nucleotide signatures such as tetranucleotide frequencies, which has already been proven most error prone during our benchmarking tests (Figure 4). Neither Gunc, nor MDMcleaner supported high contamination fractions in these genomes (Supplementary Table S1), indicating that MAGpurify may be more prone to false positives on actual genome data than our benchmarks originally suggested. This assumption may be supported by the fact that MAGpurify showed the lowest overlap with the other two assessment tools in the proportion of genomes reported as 'contaminated', especially in the case of the generally more incomplete SAGs (Supplementary Figure S4B). Gunc, on the other hand, represented the highest interquartile range of detected contaminant fractions. It also achieved the broadest outlier range when including genomes that nonetheless passed the Gunc assessment with CSS values <0.45 (Supplementary Figure S4A), meaning that, due to the way Gunc assesses apparent taxonomic conflicts within genomes (24), it is possible for a genome to have an estimated 'contamination' fraction of >80% but still be considered uncontaminated, depending on the reported CSS value. This illustrates that Gunc is well suited for genome assessments, but not for filtering/removal of respective individual contaminating contigs. Gunc generally reported by far the fewest genomes as contaminated,

Downloaded from https://academic.oup.com/nar/article/50/13/e76/6583244 by KIT Library user on 23 December 2022

while MDMcleaner reported the most (Supplementary Figure S4B). MDMcleaner consistently showed large overlaps to the other assessment tools, for example, with the majority of Gunc assessments consistently also being backed by MDMcleaner results, indicating a high reliability.

How affected are reference datasets?

Most reference database ambiguities reported by MDMcleaner among the analysed MAGs and SAGs represent potential reference database contaminations for which additional evaluations by the user are either required or at least recommended (Figure 5C). The majority of these were based on conflicting taxonomies at domain or phylum level, with average alignment identities below the genus or family cut-offs. They were therefore considered 'moderate' and 'low' indications for contaminants, respectively. While it is recommended to verify these contigs by independent analyses, it may still be justified to include such contigs in the genome submission if high genome plasticity or low contamination rates are expected. A considerable fraction of ambiguities, however, represented taxonomic conflicts at domain or phylum level with average alignment identities above genus level. Such cases demand a 'high urgency' of close evaluation by the user. Until such a contaminant can be traced exclusively to misattributed reference contigs and the correct assignment could be verified, such contigs should not be included in genome submissions. The question whether a specific contig yielding a reference database ambiguity represents a contamination only in the reference database, or also in an analysed genome, can be verified by re-aligning against different independent reference databases (e.g. GTDB, NCBI nr & NCBI RefSeq) and cross-examining the results. An example of easily verified reference database contaminations are misattributed 16S rRNA genes. However, our analyses show that genomes are not routinely scanned for these most obvious of phylogenetic markers, even within curated databases (Supplementary Table S7, Supplementary Figure S4). However, numerous unambiguous examples of reference database contaminations exist also on protein level, for example the presence of eukaryotic contigs (i.e. NCBI acc. no. NZ_LZSC01000001, NZ_AAXW01000147 and NZ_AMRJ01000061.1) in reference genomes GCF_001665235.1, GCF_000169335.1 and GCF_000300995.1) (Supplementary Table S7), which are likely misattributed due to error propagation caused by a hypothetical 'endonuclease' or 'DUF175' gene encoded on these contigs, homologues of which are mostly found in draft genomes of organisms obtained from human clinical samples. Incidentally, these three specific examples were also recently reported as eukaryotic contaminants by Steinegger *et al.* (28).

Based on these observations, we were able to compile a list of 865 contaminant contigs within GTDB reference datasets which can be provided as a 'blacklist' to future MDMcleaner runs to prevent further misattributions. This list is hosted within the public MDMcleaner repository and is expected to be extended with ongoing usage of this pipeline. We encourage users of our workflow to contribute further additions to this list, obtained from individual MDMcleaner runs on different subject genomes, in the form of issues or pull requests to the MDMcleaner repository. This way, the revision and refinement of the involved reference databases is effectively crowd-sourced, preventing continuous database contaminations and actually allowing an increasingly clear view on MDM genomes with every new analysed MAG or SAG.

The consequences: a proposal for revised MIMAG/MIMAG genome reporting standards

The herein presented MDMcleaner assessments confirm suspicions that publicly available MAG and SAG datasets of uncultured microbes are not always necessarily reliable references, even if regarded more or less contamination free by common screening procedures. However, all of the genomes identified as contaminated by MDMcleaner had been processed and published according to current best practice standards. Therefore, we want to use these examples to emphasize and discuss the need to revise these standards as they do not seem to be universally effective, especially for underrepresented members of MDM. This is being viewed as an increasing problem due to gradually increasing reference database corruption (27,28,50). The MDMcleaner workflow is a suitable countermeasure due to its high sensitivity for contaminant contigs even at small fragment sizes and its resilience against error propagation from contaminated reference database entries. Despite the broad range of outliers in the reported contamination fractions, cultivation-independent sequencing approaches can nevertheless be seen as effective sources for accurate genome information, as long as reference database error propagation can be further minimized through strict and thorough contamination screening. Correspondingly, the average decrease in genome length after filtering with MDMcleaner appears to be minimal, but the impact on the overall datasets is nonetheless huge as potential contaminations were removed from far more than half of the MAGs and almost half of the SAGs (Figure 6A and B). This indicates that MDMcleaner can provide a substantial improvement of overall genome quality for average genome completeness.

Another recently proposed countermeasure was the general exclusion of short contigs below 1.5 kb (27). This might indeed reduce the number of undetected contaminations simply by increasing the likelihood of finding conserved marker genes on the remaining contigs. However, this would nevertheless still cause a problematic fraction of misclassified contigs to remain undetected using classical contamination filtering approaches (Figures 1 and 2), considering that contigs marked as contaminants frequently exceeded 5 kb even at such robustly assignable taxonomic ranks such as domain and phylum (Figure 6C). Given the generally low loss of data but contrastingly large improvement impact of filtering, the here presented MDMcleaner workflow is a preferable alternative. Nonetheless, strict contamination filtering should apply even in cases where genome completeness would be negatively impacted. Of course, this would likely also exclude many traces of actual HGT events from the resulting genomes. However, genomes obtained exclusively from current cultivation independent methods would not be the correct basis for analysis of recent HGT events

Downloaded from https://academic.oup.com/nar/article/50/13/e76/6583244 by KIT Library user on 23 December 2022

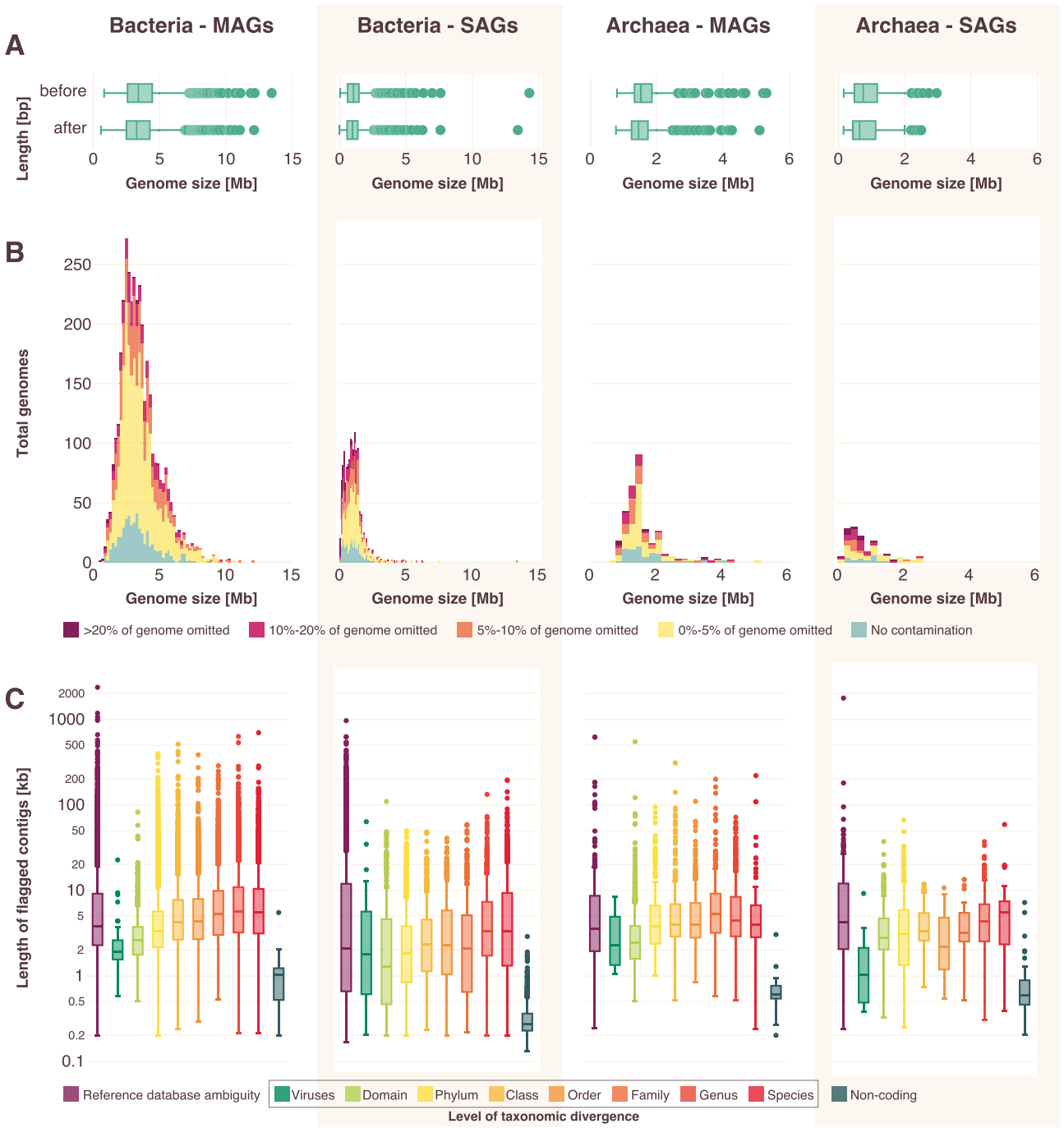

Figure 6. Impact of contig filtering by MDMcleaner. (**A**) Comparison of genome size distributions before and after MDMcleaner filtering. By applying MDMcleaner, no drastic reduction in average genome size is observed, except for a few extreme outliers. (**B**) Stacked histograms showing the extent of MDMcleaner filtering on the analysed genomes. The number of genomes of different sizes are indicated by bar-heights, while the respective fractions of contaminants filtered from the corresponding genomes are indicated by colour. Despite the low overall reduction of sequence data by MDMcleaner, the vast majority of genomes have been affected at least in a small way. (**C**) Size distribution of contigs classified as contaminants or reference ambiguities. In all cases, contaminant contigs frequently exceeded 1 kb, and often even 5 kb, demonstrating that contig size cut-offs do not sufficiently exclude contaminations.

Downloaded from https://academic.oup.com/nar/article/50/13/e76/6583244 by KIT Library user on 23 December 2022

Table 2. Summary of major findings

Contaminations are widespread among even 'high quality' public microbial dark matter genomes, with SAGs being less affected than MAGs
potential contaminations were reported in 70% of analysed MAGs ('high quality')
Highest contamination observed in analyzed MAGS: 47% ('high quality')
potential contaminations were reported in 40% of analyzed SAGs ('low contamination' and 'high quality')
Highest contamination observed in analyzed SAGs: 50% (overall 'low-contamination') & 3% ('high quality')
Even ribosomal rRNA sequences frequently overlooked by current contamination screening methods
At least 178 ribosomal RNA sequences were misattributed on phylum level or above in representative genomes of public reference datasets
381 cases of misattributed rRNA genes were found in analysed public MAGs and SAGs alone
Current genome quality assurance standards are not sufficient to tackle these problems
checkM assessments are increasingly inaccurate with decreasing genome completeness
Common nucleotide signature based approaches are suboptimal for fragmented genomes
Current MIMAG/MISAG standards emphasize quality over purity
MDMcleaner approach helps minimize contamination in microbial dark matter genomes.
Among compared screening tools, MDMcleaner was best suited for filtering contaminations in highly fragmented genomes (i.e. most MAGs and SAGs)
MDMcleaner could detect a substantial amount of contaminations in the analysed datasets, without greatly impacting average genome completeness
MDMcleaner detects and blacklists reference database contaminations, thereby preventing error propagation

anyway. Genome fragments transferred via HGT events that happened so recently, that the sequence composition has not yet adapted to the new host sufficiently enough to distinguish it from potential contamination, require an additional degree of confidence that currently is only provided by genomes from cultured isolates.

Our assessment of current publicly deposited MAGs and SAGs indicates the presence of contaminating contigs in a significant number of 'high quality' genomes, some of which even passed GTDB quality checks and were regarded as representative genomes for reference database purposes. The most striking cases are numerous misattributed SSU rRNA gene sequences representing cross-domain contaminations (Supplementary Figure S4). We therefore strongly propose a reassessment of the current definitions of MAG and SAG quality standards. Since genomic data gathered *via* cultivation-independent approaches are not directly and objectively verifiable, a higher emphasis needs to be placed on purity rather than completeness. Due to the risk of error propagation, for example, through incorrect conclusions based on best blast hit analyses, even low fractions of avoidable contamination need to be excluded.

However, with the current genome reporting standards, researchers are being motivated to submit as complete genomes as possible in order to ensure a 'high quality' status. In contrast, the labels 'moderate' and 'low quality' imply a lower scientific value. Unfortunately, the removal of potential contaminants is likely to affect completeness estimations more than, for example, the checkM marker gene multiplicities generally interpreted as contamination. This can give the false impression that decontamination procedures can be counter-productive if these lead to a downgrade in MIMAG/MISAG genome 'quality' despite the overall genome accuracy actually being improved.

In this regard, a '90% complete' genome that contains contamination, should be seen as counter-productive for reference database purposes, even if the contaminating fraction is less than 5%. Instead, genomes that appear to be free from potential contamination should be prioritized, even if they show lower completeness. Moreover, since marker gene multiplicity appears to be a less effective proxy for contamination estimation at lower completeness levels (16,51) (Supplementary Tables S1 and S2), stricter contamination cut-offs should apply for less complete genomes. Unfortunately, the opposite is enforced with the current MIMAG and MISAG standards, where higher contamination estimates are being tolerated at lower genome completeness levels, as represented by the 'medium-quality' and 'low-quality' genome standards (21). Addressing this this issue requires the application of additional screening and filtering procedures that perform direct contamination detection such as MDMcleaner, in addition to the more indirect estimations already in place.

Consequently, we propose to replace the misleading term 'quality' with two separate terms reporting 'completeness' and 'contamination' individually. The current MIMAG/MISAG cut-offs appear suitable for reporting completeness values, but in the case of contamination the stricter cut-offs of $<1\%$, $<5\%$ and $\geq 5\%$ for reporting 'low', 'moderate' and 'high' contamination, respectively, may be more adequate. In order to reflect that contamination values may be subjective, depending on the exact applied method, contamination should optimally be assessed using multiple independent screening procedures and reported as a range, for example, 'low to moderate contamination'. Furthermore, the marker gene duplicity-based contamination estimation procedure of checkM should not be applied if genome completeness estimations are below the 'high completeness' cut-off of 90%.

CONCLUSION

While contaminations are indeed prevalent in publicly submitted microbial dark matter genome datasets, cultivation-independent genome sequencing methods still remain an indispensable tool for investigating uncultivated organisms, as long as contaminations are minimized, and error propagation can effectively be prevented. However, our analyses show that current approaches are not sufficient to address this problem (Table 2), especially in the case of underrepresented and/or uncultured taxa (the major use case for metagenomics and single-cell genomics). Common pitfalls

Downloaded from https://academic.oup.com/nar/article/50/13/e76/6583244 by KIT Library user on 23 December 2022

that may have led to these short-comings include the applied method of contig classification, the selection of analysed marker genes and the underlying reference database as well as the application of nucleotide signature based approaches. A detailed description and discussion of these pitfalls is provided in Supplementary Data S1 and Supplementary Tables S1–S9, in the hope that this may help improve future contamination screening methods. We here present an easy to implement alternative workflow for detection and filtering of potential contaminants in partial and fragmented MAGs and SAGs. This approach also effectively crowdsources the detection of reference database contamination to individual researchers applying the pipeline to different novel genomes, allowing a continuous curation of the reference database and thereby preventing gradual database corruption through error propagation. This will prevent residual contaminants in submitted genomes from gradually obfuscating our view on microbial dark matter and ensure an increasingly clearer view with every new analysed and submitted MAG or SAG.

DATA AVAILABILITY

The python implementation of the MDMcleaner workflow is available at GitLab under https://github.com/KIT-IBG-5/mdmcleaner and is being distributed under a GNU general public licence v.3.0. The version used during the preparation of this publication is v0.6.0. A list of currently verified reference database contaminations, which can be passed to MDMcleaner runs in order to avoid corresponding misclassifications and error propagation, is also provided and maintained in this repository. Users are encouraged to submit additions to this blacklist, determined during individual analyses, via this repository. The reference database used by MDMcleaner expands with growing numbers of high-quality SAGs and MAGs submitted to public databases. The exact version used throughout this manuscript, and the here used benchmarking datasets, are provided at Zenodo.org under 10.5281/zenodo.5698995 and 10.5281/zenodo.5698732, respectively.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors acknowledge the support by the state of Baden-Württemberg through bwHPC and the support by the KIT-Publication Fund of the Karlsruhe Institute of Technology. The authors also want to thank Morgan Sobol for her help and for proof-reading this manuscript.

FUNDING

German Research Foundation (DFG) [320579085]. Funding for open access charge: Institute funds.
Conflict of interest statement. None declared.

REFERENCES

1. Bernard,G., Pathmanathan,J.S., Lannes,R., Lopez,P. and Bapteste,E. (2018) Microbial dark matter investigations: how microbial studies transform biological knowledge and empirically sketch a logic of scientific discovery. *Genome Biol. Evol.*, **10**, 707–715.
2. Dam,H.T., Vollmers,J., Sobol,M.S., Cabezas,A. and Kaster,A.-K. (2020) Targeted cell sorting combined with single cell genomics captures low abundant microbial dark matter with higher sensitivity than metagenomics. *Front. Microbiol.*, **11**, 1377.
3. Rinke,C., Schwientek,P., Sczyrba,A., Ivanova,N.N., Anderson,I.J., Cheng,J.-F., Darling,A., Malfatti,S., Swan,B.K., Gies,E.A. *et al.* (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, **499**, 431–437.
4. Kaster,A.K. and Sobol,M.S. (2020) Microbial single-cell omics: the crux of the matter. *Appl. Microbiol. Biotechnol.*, **104**, 8209–8220.
5. Pratscher,J., Vollmers,J., Wiegand,S., Dumont,M.G. and Kaster,A.-K. (2018) Unravelling the identity, metabolic potential and global biogeography of the atmospheric methane-oxidizing upland soil cluster α. *Environ. Microbiol.*, **20**, 1016–1029.
6. Brown,C.T., Hug,L.A., Thomas,B.C., Sharon,I., Castelle,C.J., Singh,A., Wilkins,M.J., Wrighton,K.C., Williams,K.H. and Banfield,J.F. (2015) Unusual biology across a group comprising more than 15% of domain bacteria. *Nature*, **523**, 208–211.
7. Vollmers,J., Wiegand,S. and Kaster,A.-K. (2017) Comparing and evaluating metagenome assembly tools from a microbiologist's perspective - not only size matters!*PLoS One*, **12**, e0169662.
8. Wiegand,S., Dam,H.T., Riba,J., Vollmers,J. and Kaster,A.K. (2021) Printing microbial dark matter: using single cell dispensing and genomics to investigate the patescibacteria/candidate phyla radiation. *Front. Microbiol.*, **12**, 1512.
9. Parks,D.H., Rinke,C., Chuvochina,M., Chaumeil,P.-A., Woodcroft,B.J., Evans,P.N., Hugenholtz,P. and Tyson,G.W. (2017) Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.*, **2**, 1533–1542.
10. Nayfach,S., Roux,S., Seshadri,R., Udwary,D., Varghese,N., Schulz,F., Wu,D., Paez-Espino,D., Chen,I.-M., Huntemann,M. *et al.* (2020) A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.*, **39**, 499–509.
11. Sangwan,N., Xia,F., Gilbert,J.A., Ramette,A., Tiedje,J., Tyson,G., Chapman,J., Hugenholtz,P., Allen,E., Ram,R. *et al.* (2016) Recovering complete and draft population genomes from metagenome datasets. *Microbiome*, **4**, 8.
12. Xu,Y. and Zhao,F. (2018) Single-cell metagenomics: challenges and applications. *Protein Cell*, **9**, 501–510.
13. Beaulaurier,J., Zhu,S., Deikus,G., Mogno,I., Zhang,X.S., Davis-Richardson,A., Canepa,R., Triplett,E.W., Faith,J.J., Sebra,R. *et al.* (2017) Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. *Nat. Biotechnol.*, **36**, 61–69.
14. Maguire,F., Jia,B., Gray,K.L., Lau,W.Y.V., Beiko,R.G. and Brinkman,F.S.L. (2020) Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic Islands. *Microb. Genomics*, **6**, mgen000436.
15. Sedlar,K., Kupkova,K. and Provaznik,I. (2017) Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. *Comput. Struct. Biotechnol. J.*, **15**, 48–55.
16. Becraft,E.D., Woyke,T., Jarett,J., Ivanova,N., Godoy-Vitorino,F., Poulton,N., Brown,J.M., Brown,J., Lau,M.C.Y., Onstott,T. *et al.* (2017) Rokubacteria: genomic giants among the uncultured bacterial phyla. *Front. Microbiol.*, **8**, 2264.
17. Woyke,T., Sczyrba,A., Lee,J., Rinke,C., Tighe,D., Clingenpeel,S., Malmstrom,R., Stepanauskas,R. and Cheng,J.-F. (2011) Decontamination of MDA Reagents for Single Cell Whole Genome Amplification. *PLoS One*, **6**, e26161.
18. Kogawa,M., Hosokawa,M., Nishikawa,Y., Mori,K. and Takeyama,H. (2018) Obtaining high-quality draft genomes from uncultured microbes by cleaning and co-assembly of single-cell amplified genomes. *Sci. Rep.*, **8**, 2059.
19. Parks,D.H., Imelfort,M., Skennerton,C.T., Hugenholtz,P. and Tyson,G.W. (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.*, **25**, 1043–1055.

Downloaded from https://academic.oup.com/nar/article/50/13/e76/6583244 by KIT Library user on 23 December 2022

20. Waterhouse,R.M., Seppey,M., Simão,F.A., Manni,M., Ioannidis,P., Klioutchnikov,G., Kriventseva,E. V and Zdobnov,E.M. (2018) BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.*, **35**, 543–548.
21. Bowers,R.M., Kyrpides,N.C., Stepanauskas,R., Harmon-Smith,M., Doud,D., Reddy,T.B.K., Schulz,F., Jarett,J., Rivers,A.R., Eloe-Fadrosh,E.A. *et al.* (2017) Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.*, **35**, 725–731.
22. Vollmers,J., Frentrup,M., Rast,P., Jogler,C. and Kaster,A.-K. (2017) Untangling genomes of novel Planctomycetal and Verrucomicrobial species from monterey bay kelp forest metagenomes by refined binning. *Front. Microbiol.*, **8**, 472.
23. Tennessen,K., Andersen,E., Clingenpeel,S., Rinke,C., Lundberg,D.S., Han,J., Dangl,J.L., Ivanova,N., Woyke,T., Kyrpides,N. *et al.* (2015) ProDeGe: a computational protocol for fully automated decontamination of genomes. *ISME J.*, **10**, 269–272.
24. Orakov,A., Fullam,A., Coelho,L.P., Khedkar,S., Szklarczyk,D., Mende,D.R., Schmidt,T.S.B. and Bork,P. (2021) GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome Biol.*, **22**, 178.
25. Nayfach,S., Shi,Z.J., Seshadri,R., Pollard,K.S. and Kyrpides,N.C. (2019) New insights from uncultivated genomes of the global human gut microbiome. *Nature*, **568**, 505–510.
26. Truong,D.T., Franzosa,E.A., Tickle,T.L., Scholz,M., Weingart,G., Pasolli,E., Tett,A., Huttenhower,C. and Segata,N. (2015) MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods*, **12**, 902–903.
27. Arkhipova,I.R. (2020) Metagenome proteins and database contamination. *mSphere*, **5**, e00854-20.
28. Steinegger,M. and Salzberg,S.L. (2020) Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. *Genome Biol.*, **21**, 115.
29. Lane,N. (2011) Energetics and genetics across the prokaryote-eukaryote divide. *Biol. Direct*, **6**, 35.
30. Lynch,M. and Conery,J.S. (2003) The origins of genome complexity. *Science*, **302**, 1401–1404.
31. Mira,A., Ochman,H. and Moran,N.A. (2001) Deletional bias and the evolution of bacterial genomes. *Trends Genet.*, **17**, 589–596.
32. Schloss,P.D. and Handelsman,J. (2005) Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl. Environ. Microbiol.*, **71**, 1501–1506.
33. Johnson,J.S., Spakowicz,D.J., Hong,B.Y., Petersen,L.M., Demkowicz,P., Chen,L., Leopold,S.R., Hanson,B.M., Agresta,H.O., Gerstein,M. *et al.* (2019) Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat. Commun.*, **10**, 5029.
34. Luo,C., Rodriguez-R,L.M. and Konstantinidis,K.T. (2014) MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. *Nucleic Acids Res.*, **42**, e73–e73.
35. Rodríguez-R,L. and Konstantinidis,K. (2014) Bypassing cultivation to identify bacterial species: culture-independent genomic approaches identify credibly distinct clusters, avoid cultivation bias, and provide true insights into microbial species. *Microbe Mag.*, **9**, 111–118.
36. Marine,R., McCarren,C., Vorrasane,V., Nasko,D., Crowgey,E., Polson,S.W. and Wommack,K.E. (2014) Caught in the middle with multiple displacement amplification: the myth of pooling for avoiding multiple displacement amplification bias in a metagenome. *Microbiome*, **2**, 3.
37. Solden,L., Lloyd,K. and Wrighton,K. (2016) The bright side of microbial dark matter: lessons learned from the uncultivated majority. *Curr. Opin. Microbiol.*, **31**, 217–226.
38. Geesink,P., Wegner,C.E., Probst,A.J., Herrmann,M., Dam,H.T., Kaster,A.K. and Küsel,K. (2020) Genome-inferred spatio-temporal resolution of an uncultivated Roizmanbacterium reveals its ecological preferences in groundwater. *Environ. Microbiol.*, **22**, 726–737.
39. Pratscher,J., Vollmers,J., Wiegand,S., Dumont,M.G. and Kaster,A.-K. (2018) Unravelling the identity, metabolic potential and global biogeography of the atmospheric methane-oxidizing upland soil cluster α. *Environ. Microbiol.*, **20**, 1016–1029.
40. Rinke,C., Chuvochina,M., Mussig,A.J., Chaumeil,P.-A., Davin,A.A., Waite,D.W., Whitman,W.B., Parks,D.H. and Hugenholtz,P. (2021) Resolving widespread incomplete and uneven archaeal classifications based on a rank-normalized genome-based taxonomy. bioRxiv doi: https://doi.org/10.1101/2020.03.01.972265, 17 February 2021, preprint: not peer reviewed.
41. Truong,D.T., Franzosa,E.A., Tickle,T.L., Scholz,M., Weingart,G., Pasolli,E., Tett,A., Huttenhower,C. and Segata,N. (2015) MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods*, **12**, 902–903.
42. Katayama,T., Nobu,M.K., Kusada,H., Meng,X.-Y., Hosogi,N., Uematsu,K., Yoshioka,H., Kamagata,Y. and Tamaki,H. (2020) Isolation of a member of the candidate phylum 'Atribacteria' reveals a unique cell membrane structure. *Nat. Commun.*, **11**, 6381.
43. Soucy,S.M., Huang,J. and Gogarten,J.P. (2015) Horizontal gene transfer: building the web of life. *Nat. Rev. Genet.*, **16**, 472–482.
44. Breitwieser,F.P., Pertea,M., Zimin,A. V and Salzberg,S.L. (2019) Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome Res.*, **29**, 954–960.
45. Karlicki,M., Antonowicz,S. and Karnkowska,A. (2022) Tiara: deep learning-based classification system for eukaryotic sequences. *Bioinformatics*, **38**, 344–350.
46. West,P.T., Probst,A.J., Grigoriev,I. V, Thomas,B.C., Banfield,J.F. and Banfield,J. (2018) Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Res.*, **28**, 569–580.
47. Pronk,L.J.U. and Medema,M.H. (2021) Whokaryote: distinguishing eukaryotic and prokaryotic contigs in metagenomes based on gene structure. bioRxiv doi: https://doi.org/10.1101/2021.11.15.468626, 17 November 2021, preprint: not peer reviewed.
48. Binga,E.K., Lasken,R.S. and Neufeld,J.D. (2008) Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology. *ISME J.*, **2**, 233–241.
49. Robbins,S.J., Singleton,C.M., Chan,C.X., Messer,L.F., Geers,A.U., Ying,H., Baker,A., Bell,S.C., Morrow,K.M., Ragan,M.A. *et al.* (2019) A genomic view of the reef-building coral Porites lutea and its microbial symbionts. *Nat. Microbiol.*, **4**, 2090–2100.
50. Breitwieser,F.P., Lu,J. and Salzberg,S.L. (2019) A review of methods and databases for metagenomic classification and assembly. *Brief. Bioinform.*, **20**, 1125–1136.
51. Chen,L.X., Anantharaman,K., Shaiber,A., Murat Eren,A. and Banfield,J.F. (2020) Accurate and complete genomes from metagenomes. *Genome Res.*, **30**, 315–333.

Downloaded from https://academic.oup.com/nar/article/50/13/e76/6583244 by KIT Library user on 23 December 2022