Cultivation-independent high-quality microbial genome reconstruction from environmental samples with midi-metagenomics

John Vollmers, Maximiano Correa Cassal and Anne-Kristin Kaster

Genome Res. published online May 15, 2026
Access the most recent version at doi:10.1101/gr.280099.124

Cultivation-independent high-quality microbial genome reconstruction from environmental samples with midi-metagenomics.

John Vollmers[1]*, Maximiano Correa Cassal[1] and Anne-Kristin Kaster[1,2]*

[1]Institut für Biologische Grenzflächen 5, Karlsruhe Institute of Technology, 76344 Eggenstein-Leopoldshafen

[2] Institute for Applied Biosciences, Karlsruhe Institute of Technology, 76131 Karlsruhe

*Correspondence should be addressed to john.vollmers@kit.edu and kaster@kit.edu

Abstract

Since the majority of microbial organisms still evade cultivation attempts, genomic insights into many taxa are limited to cultivation-independent approaches. However, current methods of metagenomics and single-cell genome sequencing have individual drawbacks, which can limit the quality as well as completeness of the reconstructed genomes. Current attempts to combine both approaches still use whole genome amplification techniques which are prone to bias. Here, we propose a novel approach for the purpose of genome reconstructions that utilizes the potential of cell sorting for targeted enrichment and depletion of different cell types to create distinct cell fractions with sufficient DNA amounts, circumventing amplification. By distributing sequencing efforts over these fractions as well as the original sample, co-assemblies become highly optimized for co-abundance variation based binning approaches. "Midi-metagenomics" enables accurate metagenome-assembled genome (MAG) reconstruction from individual sorted samples with higher quality than co-assembly and binning of multiple distinct samples and therefore improves analyses of uncultivated microorganisms.

Introduction

The vast majority of prokaryotes still evade cultivation attempts under laboratory conditions and therefore cannot be subjected to direct analysis *via* classical culture-based microbiological and biochemical methods (Bodor et al. 2020; Steen et al. 2019). The discrepancy between the small number of cultured prokaryotic strains compared to the vast diversity and ubiquity of uncultured microbes,commonly referred to as "microbial dark matter" (MDM; Zha et al. 2022; Rinke et al. 2013; Marcy et al. 2007; Solden et al. 2016), represents a large reservoir of biotechnological and/or pharmaceutical potential that is still untapped (Bodor et al. 2020; Escudeiro et al. 2022; Kaster and Sobol 2020)

Advances in cultivation-independent methodologies, such as metagenomics and single-cell genomics, have nowadays enabled sequence-based predictions of phylogenetic and functional characteristics of uncultivated microorganisms (Solden et al. 2016; Tyson et al. 2004; Vollmers et al. 2017) **(Fig 1),** however, with each method having distinct advantages and disadvantages. In metagenomics **(Fig 1A)**, bulk DNA from a mixed community, such as an environmental microbiome, is extracted and sequenced (Schmeisser et al. 2007). Subsequent analyses steps can then reveal the phylogenetic and functional diversity of a given community, and even enable the reconstruction of so-called "metagenome-assembled genomes" (MAGs) from uncultivated individual community members *via* contig-based binning methods (Liu et al. 2025). Unfortunately, binning of metagenomic contigs is a challenging and error-prone process, especially for highly complex communities (Ma et al. 2023) and low abundant organisms (Qayyum et al. 2025). As a result, MAGs are highly susceptible to chimerism and can show varying degrees of fragmentation and completeness in addition to contamination (Orakov et al. 2021; Vollmers et al. 2022). Furthermore, MAGs are often limited to the most abundant species present in the sample and may not reliably resolve strain variants or elements of horizontal gene transfer (Vollmers et al. 2017).

Single-cell genomics (SCG) **(Fig 1B)** circumvents this problem by directly targeting individual cells, thereby enabling reliable genome resolution on strain level (Kaster and Sobol 2020; Xu and Zhao 2018). However, since a single prokaryotic cell contains only a few femtograms of DNA, a whole genome amplification (WGA) is required, since the minimum requirement for high throughput sequencing is typically in the nanogram range (Hutchison and Venter 2006; Kalisky and Quake 2011). This is a severe disadvantage, as WGA is not only expensive and prone to contamination, but usually yields extremely uneven read coverage, constituting bias that is particularly pronounced for genomes with high GC content and usually results in single cell amplified genomes (SAGs) that are typically even more fragmented and incomplete than MAGs (Alteio et al. 2020; Kaster and Sobol 2020; Lasken and Stockwell 2007).

In order to minimize these drawbacks and maximize the advantages of both methods, there is a strong interest in combining single-cell and metagenomic approaches. A current example for such an attempt is "mini-metagenomics" (Yu et al. 2017), which targets small groups of 10-1000 cells **(Fig 1C)**. These cells are then sequenced together and subsequently treated as a simplified metagenome, to efficiently reduce random amplification bias (Alteio et al. 2020; Yu et al. 2017). The DNA yield of such a small amount of cells is, however, still not sufficient to circumvent amplification. The relatively low complexity of such mini-metagenomes should, in theory, allow for better genome reconstructions than the more complex metagenome of the original community, however, this approach is still affected by systematic WGA bias that may be caused by e.g. variations in GC content (Marine et al. 2014). Most importantly though, effective binning criteria are limited because contig abundance information is not available due to the uneven read coverage, a severe drawback that also obstructs the currently most effective binning strategy:

co-abundance variation across samples (Alneberg et al. 2014; Mattock and Watson 2023). Therefore contigs will likely have to be binned exclusively based on nucleotide signatures, which is less reliable, especially for short contigs of highly fragmented genomes (Mattock and Watson 2023; Vollmers et al. 2022).

We here present an alternative approach, termed 'midi-metagenomics', that utilizes cell sorting to create custom community fractions of sufficient cell count to circumvent the need for amplification entirely. Fluorescence-activated cell sorting (FACS) is used for targeted enrichment and depletion of different cell types to create fractions which are highly optimized for co-abundance variation based binning approaches. This way, the quality of genome reconstructions can be maximized, even if only individual samples without spatial or temporal parallels are available.

Results

Established workflow

In midi-metagenomics, the original sample population is divided into multiple fractions, in which different community members are selectively enriched or depleted *via* fluorescence-activated cell sorting (FACS) **(Fig 2A**). Possible strategies for selectively fractionating a complex community into distinct subpopulations are manifold (Sturm et al. 2023; Woyke et al. 2017). However, in this proof-of-principle study sorting was based on relatively simple gating strategies exploiting only cell characteristics easily detectable *via* FACS: cell size as determined by forward scatter and "complexity" representing cell structures and granularity as defined by side scatter gating (Lavigne et al. 1997) **(Supplemental Fig S1)**. Since soil represents one the most complex and challenging microbial communities for metagenomic analyses (Jansson and Hofmockel 2018; Vollmers et al. 2017), it was chosen as a test environment instead of controlled artificial consortia which may not faithfully represent the diversity and dynamics of natural microbiomes.

In contrast to standard single-cell and "mini-metagenomics" approaches, which require an amplification step(Rinke et al. 2013; Yu et al. 2017; Alteio et al. 2020), the midi-metagenomic methodology utilizes bulk sorts of several hundred thousand to million cells into the same fraction. Preliminary trial DNA extractions performed on bulk sorts of bacterial cultures and soil samples indicated the presence of DNA predominantly in the supernatant and not the pellet (Wiegand et al. 2021) of centrifuged cell suspensions after FACS, especially in the case of soil samples **(Supplemental Fig S2)**. This observation indicates possible cell damage caused by the sorting process, and subsequent release of cellular DNA (Binek et al. 2019; Blainey 2013; Mollet et al. 2008; Wiegand et al. 2021). Therefore, an adapted DNA extraction protocol was used for midi-metagenomic fractions, that includes an alcohol precipitation step directly from sorted cell suspensions rather than centrifuged cell pellets, thereby ensuring maximized DNA yields (**Fig 2B**). In preliminary trial runs, DNA yields ranged between 5-30 ng DNA for up to 5 million sorted cells **(Supplemental Table S1)**, which is more than sufficient for direct sequencing (Ribarska et al. 2022; Sato et al. 2019). Therefore,

based on the low input requirements of less than one nanogram (Bowman et al. 2013; Parkinson et al. 2012; Tvedte et al. 2021) for modern sequencing library preparation techniques, sorting efforts may be reduced down to 10 000 - 100 000 cells per sorted fraction in the future. The separate sequencing of genomic DNA for each fraction, as well as the original unfractionated sample **(Fig 2C)**, resulted in multiple distinct datasets for each sample.

Efficiency of cell sorting based community fractionation

The relationship between sorted fractions and corresponding unsorted samples was analyzed based on 16S rRNA gene diversity within the assembled metagenomic and midi-metagenomic fractions **(Fig 3, Supplemental Table S2)** as well as 16S rRNA amplicons with increased sequencing depth **(Supplemental Fig S3, Supplemental Table S3)**. Weighted UniFrac scores calculated from these analyses show higher beta-diversities between sorted fractions and their respective non-fractionated communities than between non-fractionated samples taken at different time points (**Fig 3)**. This increased beta-diversity represents a strong shift in relative taxon abundances within the respective microbial communities, which can be exploited for distinguishing different organisms based on differential coverage information during downstream binning attempts.

At the same time all sorted fractions show decreased alpha-diversity values, and therefore lower community complexity compared to their respective non-fractionated counterparts **(Supplemental Fig S4)**. One sample exclusively sorted on forward scatter signals (which roughly indicate cell size), clusters closer to the unsorted sample when analyzed on high-depth amplicon sequencing level **(Supplemental Fig S3)**. This indicates that the use of the forward scatter alone provides a less systematic separation of the community compared to sorting based on combined forward- and side-scatter, possibly due to the reduced resolution of morphological differences. Additional sorting metrics such as (auto)fluorescence should therefore even further improve binning efficiency.

Assembly and binning performance

Co-assemblies of standard bulk metagenomics and midi-metagenomics were compared using the same total sequencing depth of 15 Gbp (averaging at 70 million read pairs per co-assembly), equally distributed across the combined samples and fractions **(Supplemental Table S4).** Based on maximum contig length and N50 metrics, midi-metagenomic co-assemblies of sorted fractions originating from the same original sample were significantly less fragmented than co-assemblies of distinct bulk metagenomic samples **(Fig 4)**, with $p < 0.001$ as determined *via* Mann-Whitney *U* tests (MacFarland and Yates 2016).

Improved assembly metrics also affect the distribution of quality categories among the produced MAGs, as defined by the "minimum Information about a Metagenome-Assembled Genome" (MIMAG) standard, developed by the Genomic Standards Consortium (Bowers et al. 2017): Even before contamination filtering with MDMcleaner (Vollmers et al. 2022), midi-metagenomic approaches produced far more high quality MAGs (completeness >90%, contamination <5%) compared to standard metagenomic

assemblies, which predominantly consisted of only moderate quality genomes (completeness >50%, contamination <10%) or low-quality genomes (either completeness <50% or contamination >10%) with contamination values typically above 5% **(Fig 5A-C, Supplemental Tables S5 & S7)**. These trends persist across different midi-metagenomic samples and co-assembly subset groups of different sizes, despite varying community complexities, prooving the robustness of the approach.

Interestingly, midi-metagenomic MAGs represented almost twice as many distinct phyla than standard MAGs **(Fig 5D)**, illustrating another important aspect of improved assembly and binning metrics, namely improved representation of original sample diversity. This is further corroborated by detailed phylogenomic analyses of low contaminated MAGs (<5 % contamination estimate) with at least moderate (50%) completeness **(Fig 6)**, which indicate a far broader and, due to increased MAG qualities, also more reliable, phylogenomic representation by midi-metagenomic MAGs compared to standard metagenomics. In addition, in the midi-metagenomic approach a higher diversity of closely related but still distinct genomes could be reconstructed (as shown in **Fig 6** in the case of Acidobacteriota, and to a lesser extent also for Alphaproteobacteria and Actinomycetota), indicating a better resolution of sequence homologies between closely related organisms with this new method.

Comparison with alternative assembly and coverage distribution strategies

While the co-assembly of different fractions or samples enhances read coverage for assembly and co-abundance based binning, there is some debate about its applicability for standard metagenomics, as seasonal or regional variations may introduce complexities that may obstruct optimal assembly (Olm et al. 2017). In order to present an objective comparison of the performance of midi-metagenomics vs. metagenomics, we therefore tested alternative assembly strategies for the same overall sequencing depth of 15 Gbp **(Supplemental Fig S6)**. A common strategy is to perform separate single assemblies for each sample, and then map every read dataset against each individual assembly (Olm et al. 2017). This strategy greatly reduced the performance of both, metagenomic and midi-metagenomic approaches, likely due to the reduced read depth negatively affecting assembly, thereby causing an increase of “low quality” MAGs and reducing the yield of “moderate” to “high quality” MAGs **(Supplemental Fig S6A)**. Nonetheless, midi-metagenomics still performed at least as well as standard metagenomics under these conditions.

Another potential alternative sequence coverage distribution strategy is to focus sequencing and assembly efforts only on one “main” sample, and supplementing with additional lower depth “auxiliary” read datasets meant exclusively for mapping and binning purposes. However, in our trials, also this strategy could not mitigate the negative effects of the ”single assembly” strategy

for standard metagenomics, instead only showing a slightly positive effect for midi-metagenomic datasets, which thereby outperformed standard metagenomics also under these conditions **(Supplemental Fig S6B)**

Comparison with mini-metagenomics

For comparison purposes, mini-metagenomics was also applied to one sample. This approach is designed to reduce MDA bias by supplying higher amounts of input DNA by sorting and lysing multiple cells **(Fig 1C)**. Accordingly, we encountered fewer negative MDA reactions and more complete genomes using this approach compared to standard single-cell genomics. However, only two moderate quality mini-metagenomic MAGs could be recovered, both displaying high contamination estimates close to the MIMAG cut-off of 10% **(Supplemental Table S5, Supplemental Fig S6C).**

Discussion

Midi-metagenomics integrates cell sorting and metagenome sequencing approaches into a new workflow that is optimized for high quality MAG reconstruction. The key step of this approach is the seperation of the sampled microbiome into highly comparable but nevertheless distinct fractions of reduced complexity **(Fig 3, Supplemental Fig S2)** which can be sequenced directly, without involving whole genome amplification methods **(Fig 2)**.

Since the resulting fractions represent subsets of the same community at the exact same timepoint, they are optimally suited for a co-assembly strategy, maximizing the available sequencing depth for assembly as well as binning purposes. Accordingly, midi-metagenomics yielded significantly better co-assembly metrics and MAG qualities than standard metagenomics **(Figs 4 & 5),** the latter having produced significantly shorter contigs and lower quality MAGs likely due to inter-sample heterogenities which are known to generally affect co-assemblies of temporally or spatially distinct samples (Olm et al. 2017). The common alternative strategy of separately assembling each sample for metagenomics did also not improve assembly and binning outcomes (**Supplemental Fig S6**), since it severely limits the read depth available for assembly per sample (Hofmeyr et al. 2020). Compensating this effect requires substantially increased sequencing efforts and costs to reach optimal coverage for each individual sample (Ma et al. 2023; Supplemental Table S7). This limitation is especially critical given that co-abundance-based binning performs best across large numbers of parallel samples (Alneberg et al. 2014; Han et al. 2025), theoretically requiring deep sequencing of each individual sample.

Systematic comparisons with multiple independently published metagenome studies also confirm a generally lower efficiency of MAG reconstruction for metagenome approaches compared to midi-metagenomics, which yielded 411 high quality MAGs across all midi-metagenomic samples, with an average of 3 per co-assembly subgroup **(Supplemental Table S7)**. A large-scale analyses of several thousand publicly available as well as newly generated soil metagenomes by Ma et al established a direct correlation

between sequencing depth and the number and quality of generated MAGs (Ma et al. 2023). The sequencing depths in that study ranged from 0.03 to 146 Gbp, with an average depth of 14 Gbp **(Supplemental Table S7),** which is comparable to the soil metagenome control used in our study. As to be expected, the number of high quality genomes produced per metagenome varied drastically from zero to 72, averaging at 3, with larger numbers being generated at extreme sequencing depths of 100 Gbp or beyond. Interestingly, about 41% of the 3304 datasets in that study did not yield any high quality MAGs, indicating a generally relatively low efficiency of standard metagenomic approaches. Similar results were observed for two independent metagenome studies of agricultural soil samples (Hu et al. 2025; Nelkner et al. 2019; Supplemental Table S7), analyzing 8-24 samples at depths of 10-12 Gbp each. These two studies yielded considerably lower counts of high quality MAGs compared to midi-metagenomics despite surpassing the overall sequencing effort 3 to 16 fold. These comparisons also showcase that a large fraction of metagenome samples are restricted to moderate or lower quality MAGs, mostly likely due to inter-species homologies affecting assembly and binning. In contrast, midi-metagenomics yielded high quality MAGs in almost every combination of samples and fractions and in larger proportions than any of the aforementioned approaches, indicating a far more efficient distribution of sequencing and sampling efforts. The only study to reach an equivalent quality of MAGs for soil samples *via* standard metagenomics was a comparison of metagenomics and mini-metagenomics by Alteio et al (2020), comparing the genome reconstructions obtained from four deep sequenced soil metagenomes and 359 mini-metagenomes. This study, however, used considerably higher overall sequencing depth of almost 200 Gbp, averaging at ~ 50 Gb per sample, illustrating that standard metagenomics requires extreme sequencing depths to guarantee high quality genome reconstructions from soil samples, which is not financially feasible in most cases.

Our comparisons also indicate a higher effectiveness (**Supplemental Fig S6**) as well as cost-efficiency of the midi-metagenomic approach compared to mini-metagenomics, which showed at least 9x higher sequencing costs per MAG **(Supplemental Table S6)**. A major issue with mini-metagenomics is, that it targets multiple cells but still relies on whole genome amplification, which introduces bias and makes co-abundance variation based binning infeasible, thereby foregoing the main advantages of both singe cell genomics and metagenomics. These conclusions are corroborated by the results of a more thorough mini-metagenomic sampling effort by Alteio et al (2020), mentioned above. While the mini-metagenomic MAGs obtained during that study did show lower contamination values than those of standard metagenomics, they also show-cased a low overall efficiency of the approach, with few high quality MAGs being obtained from only 0.8% of the samples, despite a high combined sequencing effort of more than 190 Gbp **(Supplemental Table S7).**

The co-assembly of midi-metagenomic fractions is therefore not only the most effective, but also the most cost-efficient approach compared to current alternatives (**Supplemental Tables S6 + S7**), since sequencing efforts can be distributed across

multiple fractions for optimal binning, but can still be fully utilized for a combined assembly. The increased sequencing efficiency may be of particular interest for applications of long-read sequencing technologies, which can provide more coherent sequence context at the cost of lower read throughput (Hu et al. 2021). Furthermore, sequencing efforts do not need to be evenly distributed across midi-metagenomic fractions. Thus, the deep sequencing of an unsorted "main" fraction supplemented by auxiliary sorted fractions of lower sequencing depth can further reduce costs in cases where the research focus lies on the original community composition and the reconstruction of MAGs is only onsidered a secondary goal.

Importantly, significantly reduced contamination rates are observed in midi-metagenomic MAGs, which is especially noteworthy considering the growing complaints about reference database contaminations caused by insufficiently screened sub-quality MAGs and SAGs (Arkhipova 2020; Breitwieser et al. 2019; Vollmers et al. 2022).

We could here show that even a simple sorting set-up is sufficient for substantial improvements in both yield and quality of binned MAGs, as long as partial enrichments or depletions of different community members can be achieved. The fractionation of the sampled community can be done using FACS based on many different cell properties, of which the here utilized cell size and morphology are only the most simple examples (Sturm et al. 2023). In fact, just the act of FACS sorting itself, independent of applied criteria, already represents a general depletion of large multi-cell aggregates, extracellular DNA as well as potential stress susceptible cell types (Wiegand et al. 2021). However, more stringent sorting criteria may further improve the efficiency of the midi-metagenomic approach. Possible criteria could be labelling with Fluorescence in situ Hybridization (FISH) probes targeting 16S rRNA genes of specific taxonomic groups(Dam et al. 2020; Pratscher et al. 2018), or function based enrichments using specific gene or mRNA targeting probes(Kaster and Sobol 2020; Takahashi et al. 2020), radioactive substrate labelling (Lo et al. 2023), fluorescent labelled antibodies (Müller and Nebe-von-Caron 2010) or even just sorting based on different autofluorescence spectra caused by species specific membrane protein compositions (Kang et al. 2020).

However, it needs to be kept in mind that, due to inherient biases of the sorting process itself, i.e. the different enrichment and depletion of specific strains , any definite conclusions on relative abundances within the original community must be based on an unsorted bulk metagenome dataset. Such a bulk fraction should therefore generally be included as a reference in the fractionation workflow for each sample **(Fig 2)**. Because this unsorted bulk metagenome can seamlessly be integrated into the analyses as a dedicated fraction, this does not increase the overall sequencing efforts. Consequently, the total species richness derived by midi-metagenomics assembly can be expected to represent at least the same complexity as a standard bulk metagenome, but with the potential addition of strains that are strongly enriched in some of the sorted fractions but sequenced at depths below detection threshold in the bulk metagenome **(Figs 3 & 5D, Supplemental Fig S4, Supplemental Table S8)**. Consequently, midi-metagenomics may also serve to boost binning efforts in cases where the variability between samples may

turn out not to be sufficient for co-abundance-based binning, especially for sample locations that are hard or expensive to access for additional sampling, i.e. deep-sea sediments. The exact sorting criteria do not even need to be decided beforehand as a glycerol stock of frozen sample can be revisited for sorting after a preliminary whole-community metagenome analyses. Although the fractionation process does require access to dedicated equipment such as a FACS- or microfluidic-based cell sorter which are typically priced at around 200.000€, the MDA-free workflow allows for much more streamlined and cost-effective set-ups compared to mini-metagenomics: Without the highly contamination sensitive MDA-step, the necessity of clean-room standards is removed, greatly simplifying the required infrastructure as well as expertise and potentially even allowing to outsource the sorting process to external facilities that already possess adequate sorting-devices. Consequently, midi-metagenomics represents a novel and improved technique that is widely accessible to the scientific community and can significantly enhance the quality and reliability of prokaryotic genome reconstruction from environmental samples.

Methods

Microbial Samples

To evaluate midi-metagenomics performance compared to metagenomics, soil samples were collected at the Karlsruhe Institute of Technology (KIT) – Campus North, Eggenstein-Leopoldshafen (49°5’48.8’’N, 8°25’55.6’’E), Germany, during four different periods of time: October 7th, 2020, May 25th, 2020, August 10th, 2021, and February 15th, 2022. From each sample, several grams were directly frozen at -80 °C immediately after collection for subsequent standard metagenome DNA extraction and sequencing. Five grams of each sample was then prepared for cell sorting by adding 30 mL of filtered, autoclaved and UV-sterilized Phosphate Buffer Saline (PBS) solution, brief vortexing to disrupt aggregates and dislocate cells attached to debris, and subsequent pelleting and removal of debris by brief centrifugation at 2,000 × g. Sterile glycerol was added to a final concentration of 30% as an anti-freezing agent and the samples were stored at -80°C until further processing. An overview of all samples is given in **Table 1**.

Fluorescence-Activated Cell Sorting (FACS)

Prior to sorting, the samples aliquoted for midi-metagenomics were centrifuged for 1 min at 15,871 × g and 20 °C. The supernatant was discarded and after resuspension of the pellet in 1 mL PBS, 5 µl SYBR® Green I was added to all samples. The samples were then vortexed, incubated for 20 min at 4 °C and subsequently pelleted again by centrifugation for 1 min at 15,871 × g. Each pellet was then washed twice with 1 mL PBS.

Before loading the sample into the FACS (BD FACSMelody™, Becton, Dickinson and Company, New Jersey, USA), an unlabeled negative control was filtered into a 5 mL FACS tube using a sterile SYSMEX CellTrics® filter with 20 µM mesh size and then diluted with PBS. The negative control was used to compare the difference of fluorescence signals for a correct gating that included only

labelled cells. Subsequently, the same procedure was applied to the SYBR-labelled samples. A threshold was set up in order to disregard smaller particles such as debris during the sorting process and an excitation wavelength of 488 nM was used.

For samples “summer 21”, “winter21/22” and “winter22/23” cells were sorted into five different groups via gatings based on plotting fluorescence intensity against the Forward Scatter Signal (FSC) and Side Scatter Signal (SSC), which are roughly proportional to cell size and complexity, respectively (**Supplemental Table S1 & Supplemental Fig S1**). For sample “autumn20” only two groups were sorted, according to size measured by differences in FSC (**Supplemental Table S1**). Configurations for fluidics, optical and electronic settings were kept constant for all sorting runs, as specified in **Supplemental Table S1.** No compensation was applied, as only one fluochrome was used. After sorting, the cells were stored at -80 °C until further processing. An overview of the Fractions produced per sample is included in **Table 1**.

DNA Extraction

For metagenomics of the unsorted sample, DNA was extracted with the Dneasy PowerSoil Kit (Qiagen, Hilden, Germany) following the manufacturer’s instructions. For midi-metagenomics community fractions, DNA was extracted directly from FACS sorted cell suspensions consisting of 4×10^6 cells. First, the cells were freeze-thawed three times using liquid nitrogen and a 60 °C water bath. Then, bead beating was performed three times for 30 s at 6 m/s using one tube of lysing matrix for each fraction (Cat.#6914-800, MP Biomedicals, Ohio, USA) and an MP Bio Fast Prep®-24 homogenizer (MP Biomedicals, Ohio, USA). Beads and cell debris were pelleted by centrifugation at 14,000 ×g for 5 min and the supernatant was subjected to standard alcohol precipitation using 1 volume of 80% isopropanol, 0.1 volume 3 M Sodium Acetate and 340 µg Linear Polyacrylamide. After a subsequent wash step with ice cold 70% ethanol the resulting DNA pellet was resuspended with 100 µl PCR-grade water followed by further purification *via* solid-phase reversible immobilization using 1.5 volume of AMPure XP Beads (Beckman Coultier™) and final elution in 20 µl 1× TE. All extracted DNA was immediately stored at -20 °C until use.

Polymerase Chain Reaction (PCR) for Amplicons

Amplicon sequencing was performed using a nested PCR approach. Almost full-length PCR products were obtained in a preliminary PCR using 1.25U OneTaq® Quick-Load® DNA Polymerase (New England BioLabs, Ipswich, MA, USA), 200 µM mixed dNTPs, 500 µM biology-grade Bovine Serum Albumin (BSA) (Thermo Fisher Scientific, Waltham, Massachusetts, USA) and 0.2 µM of each universal bacterial forward and reverse primer 27F (5’-AGRGTTYGATYMTGGCTCAG-3’) and 1492R(5’-AGRGTTYGATYMTGGCTCAG-3’). PCR products were purified using DNA Clean & Concentrator™-5 columns (Zymo Research Europe GmbH, Irvine, California, USA) according to the manufacturer’s instructions. The purified product was then used as template for

a subsequent amplicon PCRs using 0.5 U Q5® High-Fidelity DNA Polymerase (New England Biolabs, Ipswich, MA, USA) 0.5 U, 200 µM dNTP Solution Mix (New England Biolabs), Q5® High GC Enhancer, 0.1 µg/µl BSA (Thermo Fisher Scientific, Waltham, Massachusetts, USA) and 0.2 µM of each universal bacterial primer 341F (5'-AGRGTTYGATYMTGGCTCAG-3') and 518R (5'-AGRGTTYGATYMTGGCTCAG-3'), targeting the V3 hypervariable region.

Sequencing

All libraries were prepared using the NEBNext® Ultra™ II FS DNA Library Prep Kit for Illumina® (New England Biolabs, Ipswich, MA, USA), according to the manufacturer's instructions. Libraries were sequenced on an Illumina NextSeq 550® (New England Biolabs, Ipswich, MA, USA) device using 300 cycles and a paired-end approach.

Read processing and assembly

Reads were quality trimmed and adapter-clipped using Trimmomatic v.0.36, bbduk v.35.69 and cutadapt v.1.14 successively (Bolger et al. 2014; Martin 2011; https://bbmap.org). Overlapping read pairs were identified and merged using FLASH v.1.2.11 (Magoč and Salzberg 2011). For amplicon datasets, reads were clustered into Amplicon Sequence Variants (ASV) using the Qiime2 pipeline with dada2 (Bolyen et al. 2019; Callahan et al. 2016). To account for different sequencing depths the produced AVS were rarefied to a value of 42000 which was determined via preliminary rarefaction analysis. Rarefied ASVs were subsequently taxonomically classified using RDP Classifier v1.24 (Wang and Cole 2024) and SINA v1.7.2 (Pruesse et al. 2012). Shotgun datasets were arranged into 81 co-assembly subgroups representing all possible subsets of three to six datasets per metagenome or midi-metagenome (**Supplemental Table S4**). In order to enable standardized assemblies with 15 Gbp total read input each, shotgun datasets were randomly subsampled down to 2.5, 3.75, 3 and 5 Gbp when possible, and coassembled at equal amounts for each co-assembly subgroup. Additional assemblies with non-uniform sequencing depth distribution were also performed using unsorted metagenome samples as "main" datasets with 12-13 Gbp sequencing depth, and 5 "auxilliary" datasets subsampled to 0.4-0.5 Gbp each. Assemblies were performed using MEGAHIT v1.2.9 (Li et al. 2015).

For 16S rRNA gene diversity analysis in shotgun assemblies, 16S rRNA genes were extracted from all assemblies using barrnap (https://github.com/tseemann/barrnap), clustered at 99% sequence identity level using VSEARCH v2.21.1 (Rognes et al. 2016), and cluster-representatives were aligned using SINA v1.7.2. Alignments were then filtered in order to retain only those that fully overlap a defined sequence window of 600 bp roughly representing the hypervariable regions V3-V6. For beta-diversity analyses, OTU tables were generated based the filtered sequences and read coverages determined via coverm v0.7.0 (Aroney et al. 2025). Beta-diversity analyses were then performed using Qiime2 and taxonomic classifications were performed as described above.

MAG reconstruction and analyses

For each co-assembly, three different binning tools were used in parallel: Metabat2 v.2.15 (Kang et al. 2019), CONCOCT (Alneberg et al. 2014) and Rosella v.0.4.1 (Newell 2023). For Midi-metagenomic approaches, Rosella was substituted for Maxbin (Wu et al. 2016), as Rosella did not function without co-abundance information and Maxbin utilizes additional taxonomic and marke-gene criteria that may optimize results for mini-metagenomic and single cell genomic assemblies that lack reliable coverage information (Kaster and Sobol 2020; Marine et al. 2014). Resulting bins were pre-assessed and filtered using MDMcleaner. Quality categories were then determined based on re-assessments using checkm2 (Chklovski et al. 2023). Taxonomic classifications were based on GTDB-TK v2.1.1 (Chaumeil et al. 2020).

dRep v.3.4.0 (Olm et al. 2017) was employed to identify groups of redundant MAGs created by different assemblies or binning tools and to select the respective most representative MAG. Similarities between MAGs were additionally determined and visualized based on gene-content as previously described elsewhere (Howat et al. 2018).

Data access

All raw sequencing data and processed dereplicated high quality MAG sequences have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA900514. The complete set of processed sequencing data, including redundant and low quality MAGs with completeness estimates of at least 50% and contamination estimates below 25% have been submitted to zenodo under DOI: 10.5281/zenodo.13150466

Competing interest statement

The authors declare that they have no competing interests

Acknowledgements

The Authors acknowledge support by the state of Baden-Württemberg through bwHPC.

This work was financially supported through the Helmholtz Association program “Materials Systems Engineering” under the topic “Adaptive and Bioinstructive Materials Systems” (project ID: 43.33.11) and by the German government, through BMBF project MicroMatrix (project ID: 161L0284A)

Author contributions were as follows: Study conception and design: John Vollmers, Anne-Kristin Kaster; data collection: Maximiano Cassal, John Vollmers, Analysis and interpretation of results: John Vollmers, manuscript preparation: John Vollmers, Maximiano Cassal, Anne-Kristin Kaster, Funding: Anne-Kristin Kaster

References

Alneberg J, Bjarnason BS, De Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. 2014. Binning metagenomic contigs by coverage and composition. *Nat Methods* **11**: 1144–1146.

Alteio LV, Schulz F, Seshadri R, Varghese N, Rodriguez-Reillo W, Ryan E, Goudeau D, Eichorst SA, Malmstrom RR, Bowers RM, et al. 2020. Complementary Metagenomic Approaches Improve Reconstruction of Microbial Diversity in a Forest Soil. *mSystems* **5**: 10.1128/msystems.00768-19.

Arkhipova IR. 2020. Metagenome Proteins and Database Contamination. *mSphere* **5**: 10.1128/msphere.00854-20.

Aroney STN, Newell RJP, Nissen JN, Camargo AP, Tyson GW, Woodcroft BJ. 2025. CoverM: read alignment statistics for metagenomics. *Bioinformatics* **41**: btaf147.

Binek A, Rojo D, Godzien J, Rupérez FJ, Nuñez V, Jorge I, Ricote M, Vázquez J, Barbas C. 2019. Flow Cytometry Has a Significant Impact on the Cellular Metabolome. *J Proteome Res* **18**: 169–181.

Blainey PC. 2013. The future is now: single-cell genomics of bacteria and archaea. *FEMS Microbiol Rev* **37**: 407–427.

Bodor A, Bounedjoum N, Vincze GE, Erdeiné Kis Á, Laczi K, Bende G, Szilágyi Á, Kovács T, Perei K, Rákhely G. 2020. Challenges of unculturable bacteria: environmental perspectives. *Rev Environ Sci Biotechnol* **19**: 1–22.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, et al. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* **37**: 852–857.

Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA, 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.

Bowman SK, Simon MD, Deaton AM, Tolstorukov M, Borowsky ML, Kingston RE. 2013. Multiplexed Illumina sequencing libraries from picogram quantities of DNA. *BMC Genomics* **14**: 466.

Breitwieser FP, Pertea M, Zimin AV, Salzberg SL. 2019. Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome Res* **29**: 954–960.

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* **13**: 581–583.

Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. 2020. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**: 1925–1927.

Chklovski A, Parks DH, Woodcroft BJ, Tyson GW. 2023. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat Methods* **20**: 1203–1212.

Dam HT, Vollmers J, Sobol MS, Cabezas A, 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.

Escudeiro P, Henry CS, Dias RPM. 2022. Functional characterization of prokaryotic dark matter: the road so far and what lies ahead. *Curr Res Microb Sci* **3**: 100159.

Gibbons JD. 2005. Median Test, Brown–Mood. In *Encyclopedia of Statistical Sciences* (eds. S. Kotz, C.B. Read, N. Balakrishnan, and B. Vidakovic), Wiley https://doi.org/10.1002/0471667196.ess0181.pub2.

Han H, Wang Z, Zhu S. 2025. Benchmarking metagenomic binning tools on real datasets across sequencing platforms and binning modes. *Nat Commun* **16**: 2865.

Hofmeyr S, Egan R, Georganas E, Copeland AC, Riley R, Clum A, Eloe-Fadrosh E, Roux S, Goltsman E, Buluç A, et al. 2020. Terabase-scale metagenome coassembly with MetaHipMer. *Sci Rep* **10**: 10689.

Howat AM, Vollmers J, Taubert M, Grob C, Dixon JL, Todd JD, Chen Y, Kaster A-K, Murrell JC. 2018. Comparative Genomics and Mutational Analysis Reveals a Novel XoxF-Utilizing Methylotroph in the Roseobacter Group Isolated From the Marine Environment. *Front Microbiol* **9**. https://doi.org/10.3389/fmicb.2018.00766.

Hu T, Chitnis N, Monos D, Dinh A. 2021. Next-generation sequencing technologies: An overview. *Hum Immunol* **82**: 801–811.

Hu X, Liu J, Liang A, Gu H, Liu Z, Jin J, Wang G. 2025. Soil metagenomics reveals reduced tillage improves soil functional profiles of carbon, nitrogen, and phosphorus cycling in bulk and rhizosphere soils. *Agric Ecosyst Environ* **379**: 109371.

Hutchison CA, Venter JC. 2006. Single-cell genomics. *Nat Biotechnol* **24**: 657–658.

Jansson JK, Hofmockel KS. 2018. The soil microbiome — from metagenomics to metaphenomics. *Curr Opin Microbiol* **43**: 162–168.

Kalisky T, Quake SR. 2011. Single-cell genomics. *Nat Methods* **8**: 311–314.

Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, Wang Z. 2019. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**: e7359.

Kang S-M, de Josselin de Jong E, Higham SM, Hope CK, Kim B-I. 2020. Fluorescence fingerprints of oral bacteria. *J Biophotonics* **13**: e201900190.

Kaster A-K, Sobol MS. 2020. Microbial single-cell omics: the crux of the matter. *Appl Microbiol Biotechnol* **104**: 8209–8220.

Lasken RS, Stockwell TB. 2007. Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnol* **7**: 19.

Lavigne S, Bossé M, Boulet LP, Laviolette M. 1997. Identification and analysis of eosinophils by flow cytometry using the depolarized side scatter-saponin method. *Cytometry* **29**: 197–203.

Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**: 1674–1676.

Liu S, Rodriguez JS, Munteanu V, Ronkowski C, Sharma NK, Alser M, Andreace F, Blekhman R, Błaszczyk D, Chikhi R, et al. 2025. Analysis of metagenomic data. *Nat Rev Methods Primer* **5**: 5.

Lo H-Y, Wink K, Nitz H, Kästner M, Belder D, Müller JA, Kaster A-K. 2023. scMAR-Seq: a novel workflow for targeted single-cell genomics of microorganisms using radioactive labeling. *mSystems* **8**: e00998-23.

Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R. 2011. UniFrac: an effective distance metric for microbial community comparison. *ISME J* **5**: 169–172.

Ma B, Lu C, Wang Y, Yu J, Zhao K, Xue R, Ren H, Lv X, Pan R, Zhang J, et al. 2023. A genomic catalogue of soil microbiomes boosts mining of biodiversity and genetic resources. *Nat Commun* **14**: 7318.

MacFarland TW, Yates JM. 2016. Mann–Whitney U Test. In *Introduction to Nonparametric Statistics for the Biological Sciences Using R*, pp. 103–132, Springer International Publishing, Cham.

Magoč T, Salzberg SL. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**: 2957–2963.

Marcy Y, Ouverney C, Bik EM, Lösekann T, Ivanova N, Martin HG, Szeto E, Platt D, Hugenholtz P, Relman DA, et al. 2007. Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc Natl Acad Sci* **104**: 11889–11894.

Marine R, McCarren C, Vorrasane V, Nasko D, Crowgey E, Polson SW, Wommack KE. 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.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**: 10–12.

Mattock J, Watson M. 2023. A comparison of single-coverage and multi-coverage metagenomic binning reveals extensive hidden contamination. *Nat Methods* **20**: 1170–1173.

Mollet M, Godoy-Silva R, Berdugo C, Chalmers JJ. 2008. Computer simulations of the energy dissipation rate in a fluorescence-activated cell sorter: Implications to cells. *Biotechnol Bioeng* **100**: 260–272.

Müller S, Nebe-von-Caron G. 2010. Functional single-cell analyses: flow cytometry and cell sorting of microbial populations and communities. *FEMS Microbiol Rev* **34**: 554–587.

Nelkner J, Henke C, Lin TW, Pätzold W, Hassa J, Jaenicke S, Grosch R, Pühler A, Sczyrba A, Schlüter A. 2019. Effect of Long-Term Farming Practices on Agricultural Soil Microbiome Members Represented by Metagenomically Assembled Genomes (MAGs) and Their Predicted Plant-Beneficial Genes. *Genes* **10**: 424.

Newell R. 2023. Bioinformatic methods for genome-centric metagenomics. Queensland University of Technology https://doi.org/10.5204/thesis.eprints.237953.

Olm MR, Brown CT, Brooks B, Banfield JF. 2017. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J* **11**: 2864–2868.

Orakov A, Fullam A, Coelho LP, Khedkar S, Szklarczyk D, Mende DR, Schmidt TSB, Bork P. 2021. GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome Biol* **22**: 178.

Parkinson NJ, Maslau S, Ferneyhough B, Zhang G, Gregory L, Buck D, Ragoussis J, Ponting CP, Fischer MD. 2012. Preparation of high-quality next-generation sequencing libraries from picogram quantities of target DNA. *Genome Res* **22**: 125–133.

Pratscher J, Vollmers J, Wiegand S, Dumont MG, 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.

Pruesse E, Peplies J, Glöckner FO. 2012. SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**: 1823–1829.

Qayyum H, Talib MS, Ali A, Kayani MUR. 2025. Evaluating the potential of assembler-binner combinations in recovering low-abundance and strain-resolved genomes from human metagenomes. *Heliyon* **11**: e41938.

Ribarska T, Bjørnstad PM, Sundaram AYM, Gilfillan GD. 2022. Optimization of enzymatic fragmentation is crucial to maximize genome coverage: a comparison of library preparation methods for Illumina sequencing. *BMC Genomics* **23**: 92.

Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, Darling A, Malfatti S, Swan BK, Gies EA, et al. 2013. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**: 431–437.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**: e2584.

Sato MP, Ogura Y, Nakamura K, Nishida R, Gotoh Y, Hayashi M, Hisatsune J, Sugai M, Takehiko I, Hayashi T. 2019. Comparison of the sequencing bias of currently available library preparation kits for Illumina sequencing of bacterial genomes and metagenomes. *DNA Res Int J Rapid Publ Rep Genes Genomes* **26**: 391–398.

Schmeisser C, Steele H, Streit WR. 2007. Metagenomics, biotechnology with non-culturable microbes. *Appl Microbiol Biotechnol* **75**: 955–962.

Solden L, Lloyd K, Wrighton K. 2016. The bright side of microbial dark matter: lessons learned from the uncultivated majority. *Curr Opin Microbiol* **31**: 217–226.

Steen AD, Crits-Christoph A, Carini P, DeAngelis KM, Fierer N, Lloyd KG, Cameron Thrash J. 2019. High proportions of bacteria and archaea across most biomes remain uncultured. *ISME J* **13**: 3126–3130.

Sturm G, Mojarrad M, Kaster A-K. 2023. Targeted Cell Labeling and Sorting of Prokaryotes for Cultivation and Omics Approaches. *Microb Physiol* **33**: 63–84.

Takahashi H, Horio K, Kato S, Kobori T, Watanabe K, Aki T, Nakashimada Y, Okamura Y. 2020. Direct detection of mRNA expression in microbial cells by fluorescence in situ hybridization using RNase H-assisted rolling circle amplification. *Sci Rep* **10**: 9588.

Tvedte ES, Michalski J, Cheng S, Patkus RS, Tallon LJ, Sadzewicz L, Bruno VM, Silva JC, Rasko DA, Dunning Hotopp JC. 2021. Evaluation of a high-throughput, cost-effective Illumina library preparation kit. *Sci Rep* **11**: 15925.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**: 37–43.

Vollmers J, Wiegand S, Kaster A-K. 2017. Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist's Perspective - Not Only Size Matters! ed. F. Rodriguez-Valera. *PLOS ONE* **12**: e0169662.

Vollmers J, Wiegand S, Lenk F, Kaster A-K. 2022. How clear is our current view on microbial dark matter? (Re-)assessing public MAG & SAG datasets with MDMcleaner. *Nucleic Acids Res* **50**: e76–e76.

Wang Q, Cole JR. 2024. Updated RDP taxonomy and RDP Classifier for more accurate taxonomic classification. *Microbiol Resour Announc* **13**: e01063-23.

Wiegand S, Dam HT, Riba J, Vollmers J, Kaster A-K. 2021. Printing Microbial Dark Matter: Using Single Cell Dispensing and Genomics to Investigate the Patescibacteria/Candidate Phyla Radiation. *Front Microbiol* **12**. https://doi.org/10.3389/fmicb.2021.635506.

Woyke T, Doud DFR, Schulz F. 2017. The trajectory of microbial single-cell sequencing. *Nat Methods* **14**: 1045–1054.

Wu Y-W, Simmons BA, Singer SW. 2016. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**: 605–607.

Xu Y, Zhao F. 2018. Single-cell metagenomics: challenges and applications. *Protein Cell* **9**: 501–510.

Yu FB, Blainey PC, Schulz F, Woyke T, Horowitz MA, Quake SR. 2017. Microfluidic-based mini-metagenomics enables discovery of novel microbial lineages from complex environmental samples. *eLife* **6**: e26580.

Zha Y, Chong H, Yang P, Ning K. 2022. Microbial Dark Matter: from Discovery to Applications. *Genomics Proteomics Bioinformatics* **20**: 867–881.

Figure Legends:

Figure 1: Current culture-independent methodologies. In (**A**) Metagenomics the entire DNA of an environmental community is sequenced. Assembled contigs are binned into metagenome-assembled genomes (MAGs). In (**B**) single-cell genomics (SCG), individual cells are isolated, sequenced and analyzed. Due to little DNA content per cell, whole genome amplification (WGA) is required. In (C) Mini-metagenomics typically pools of 5-1000 cells are sortied and sequenced. While complexity is lower than for standard metagenomics, binning is still required, and the low DNA content of small cell pools still necessitates WGA. Created with BioRender.com

Figure 2: Midi-metagenomics workflow. (A) Part of the sample community is fractionated into distinct groups of several hundred thousand to millions of cells by cell sorting. Different cell types are not separated with absolute stringency, but differentially enriched **(B)** DNA is extracted separately from each fraction, as well as the original unsorted sample. **(C)** Extracted DNA is sequenced directly without whole genome amplification (WGA). **(D)** Since the resulting read datasets represent different enrichments based on the same original community, they are optimal for co-assembly as well as co-abundance variation-based binning approaches. An unbiased representation of the source community is achieved by also including the original unsorted sample in the analyses. Created with BioRender.com

Figure 3: 16S rRNA gene based diversity among different (midi-)metagenomic fractions of different samples Clustering is based on weighted UniFrac (Lozupone et al. 2011) beta-diversity scores and is shown as cladogram on the left. The background colouring of Y-Axis labels on the right side indicates the respective origin-sample. Stacked bar charts indicate the community composition of each sample and fraction, with different phyla being indicated by a distinct colour code as indicated above the plot, and relative abundances being indicated by bar heights according to the X-axis below the plot. Sorted midi-metagenomic fractions are indicated by pictograms and abbreviations as given in the legend on the right

Figure 4: Comparison of assembly metrics for metagenomic and midi-metagenomic approaches in dependance of co-assembled samples and fractions. Scatterplots showing Metagenomic results as red, and Midi-metagenomic results as purple dots. Trendlines and corresponding confidence areas were determined by regression analysis and are indicated by dashed lines and background coloring, respectively. **(A)** Maximum contig lengths **(B)** Contig N50 values. The significance of differences in the distribution between Metagenomic and Midi-metagenomic assembly metrics were determined *via* Mann-Whitney *U* tests (MacFarland and Yates 2016). MG = Metagenome, Midi-MG = Midi-metagenome

Figure 5: Comparison of quality metrics and diversity of MAGs obtained from standard metagenomic and midi-metagenomic co-assemblies. **(A & B)** Boxplots showing the distribution of checkm2 completeness and contamination estimates, respectively. The difference between metagenomic and midi-metagenomic results was found to be statistically significant ($p < .01$ based on Moods Median Test; Gibbons 2005) in both cases. A more detailed plot showing individual results is given in **Supplemental Fig S5 (C)** Average number of MAGs belonging to different quality categories obtained by standard metagenomics and midi-metagenomics. **(D)** Relative fractions of total phylum level diversity detected in the unbinned metagenomic co-assemblies that are represented by metagenomic and midi-metagenomic MAGs, respectively. MG = Metagenome, Midi-MG = Midi-metagenome

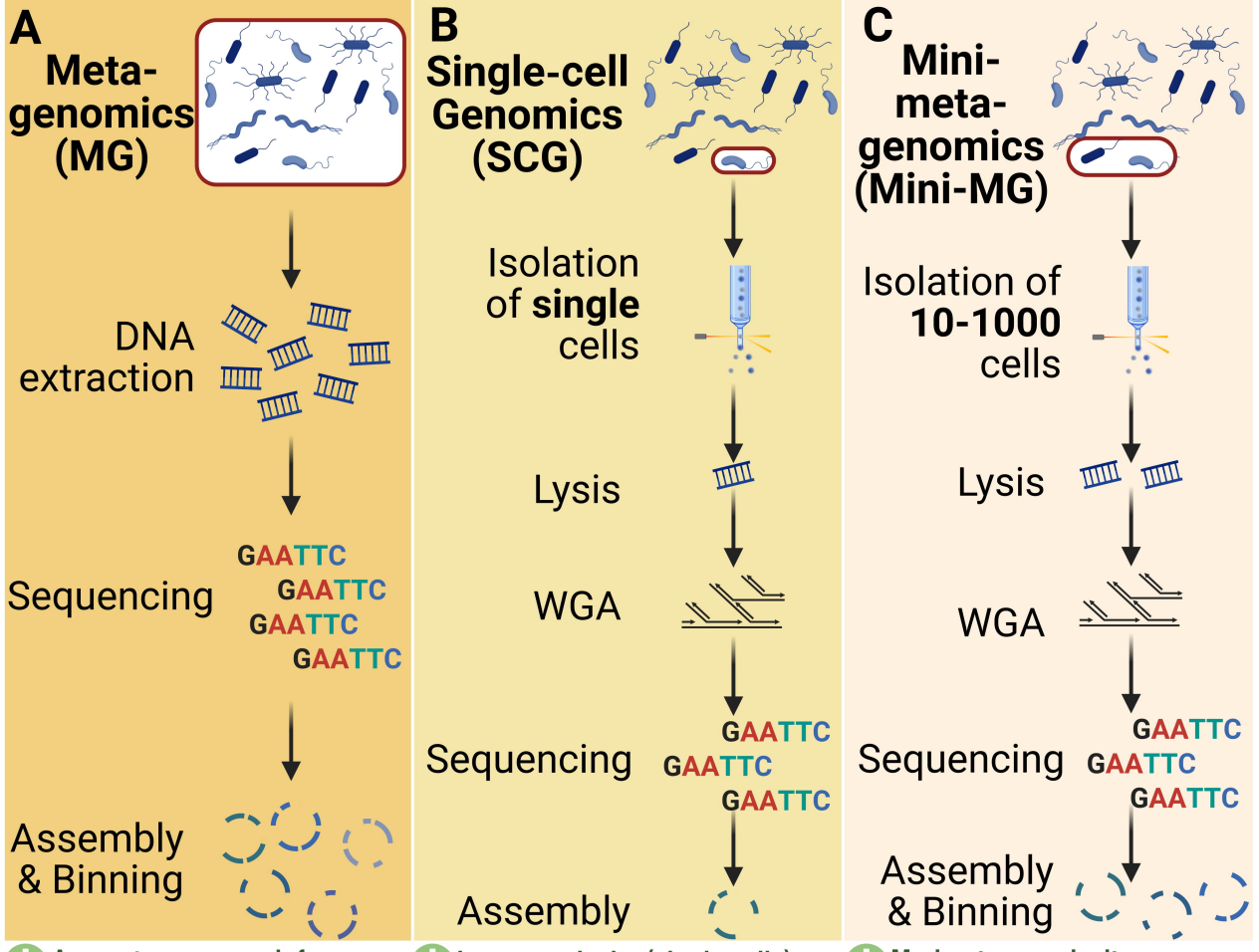
Figure 6: Multi Locus Sequence Analyses (MLSA) based phylogeny of representative MAGs and related reference genomes. Maximum likelyhood phylogenetic clustering based on 61 single copy orthologs shared by all comparison genomes, concatenated to a total length of 7178 amino acids. The software tool dRep (Olm et al. 2017) was used to group all MAGs on species level based on ANI comparisons, and to rank the members of each group based on genome quality. Only groups with representatives showing >50% completeness and <=5% contamination were considered, resulting in 30 groups labelled C1-C30, for each of which only the best representative is compared. Bubble Plots next to each group designation indicate the number and average dRep score of each group, as indicated by the legend on the upper left. α = Alphaproteobacteria, δ = Gammaproteobacterial, MG = Metagenomics, Midi-MG = Midi-metagenomics

534

535 # Tables

536 **Table 1: Overview of samples and fractions.** Fraction abbreviations: BC = “Big Complex”, MC = “Medium Complex”, SC = “Small Complex”, BNC = “Big Non-Complex”, SNC = “Small Non-Complex”,

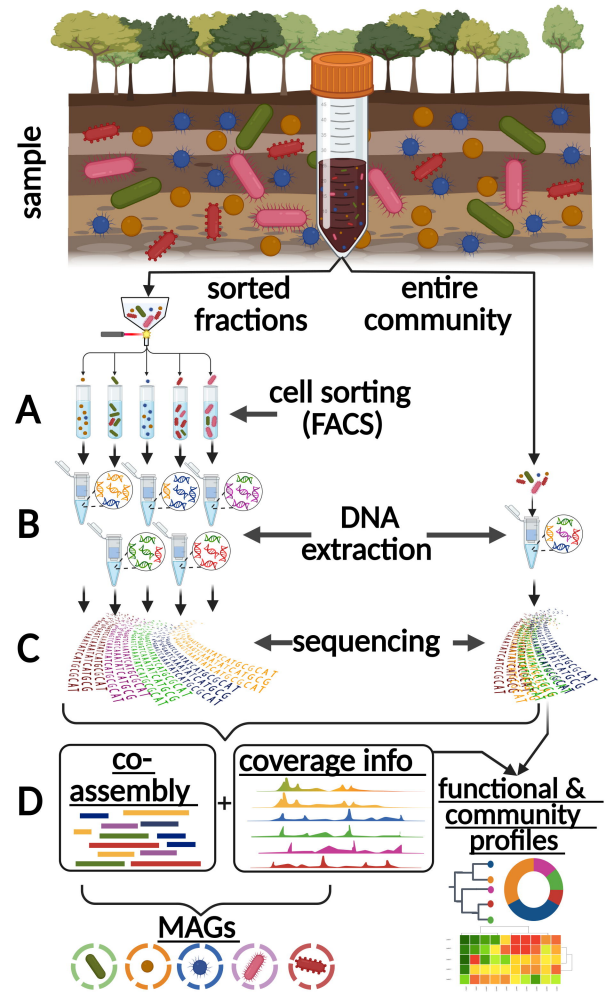
Sample	Sampling location	Sampling date	Fractions produced
spring20	KIT, Campus North 49°5’48.8’’N, 8°25’55.6’’E	15.05.2020	only unsorted
autumn20		07.10.2020	unsorted, big & small
summer21		10.08.2021	unsorted, BC, MC, SC, BNC & SNC
Winter21/22		15.02.2022	unsorted, BC, MC, SC, BNC & SNC
Autumn22		07.10.2022	only unsorted
Winter22/23			unsorted, BC, MC, SC, BNC & SNC

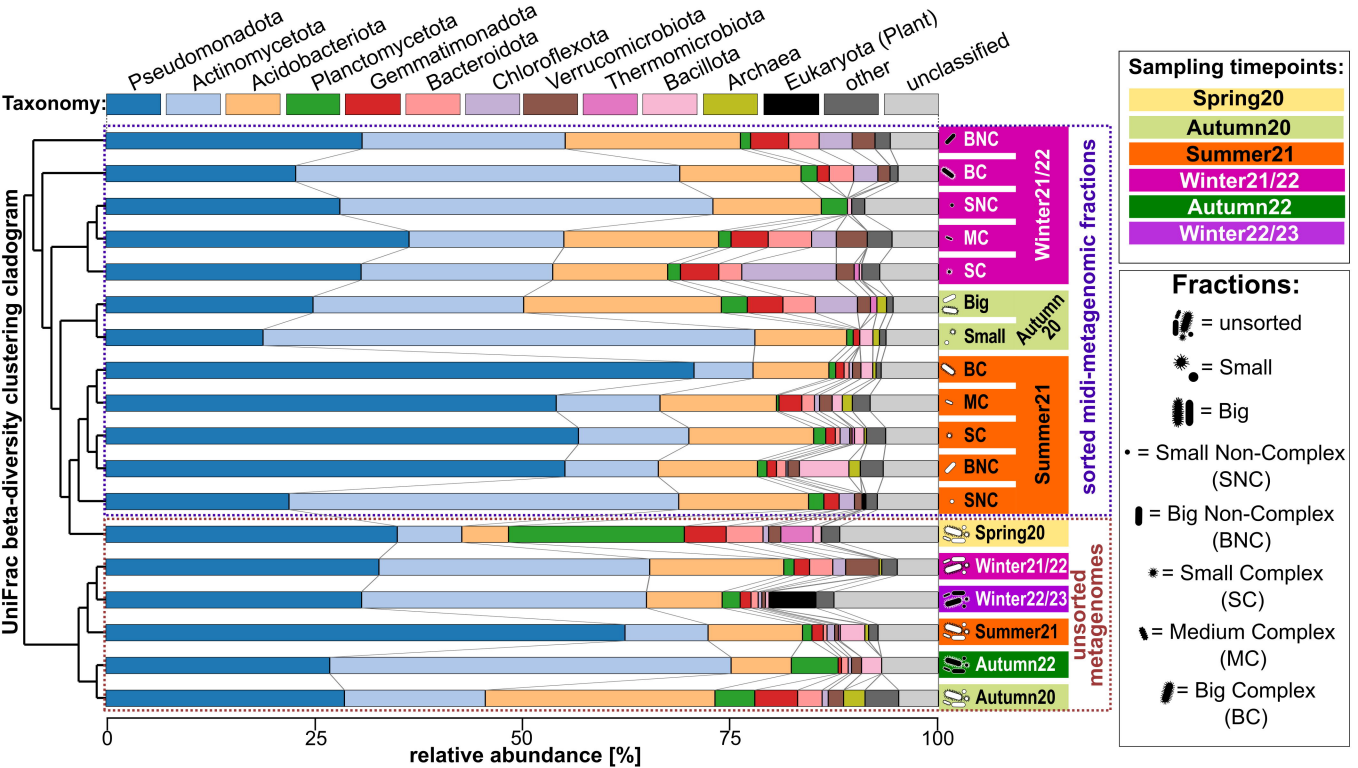


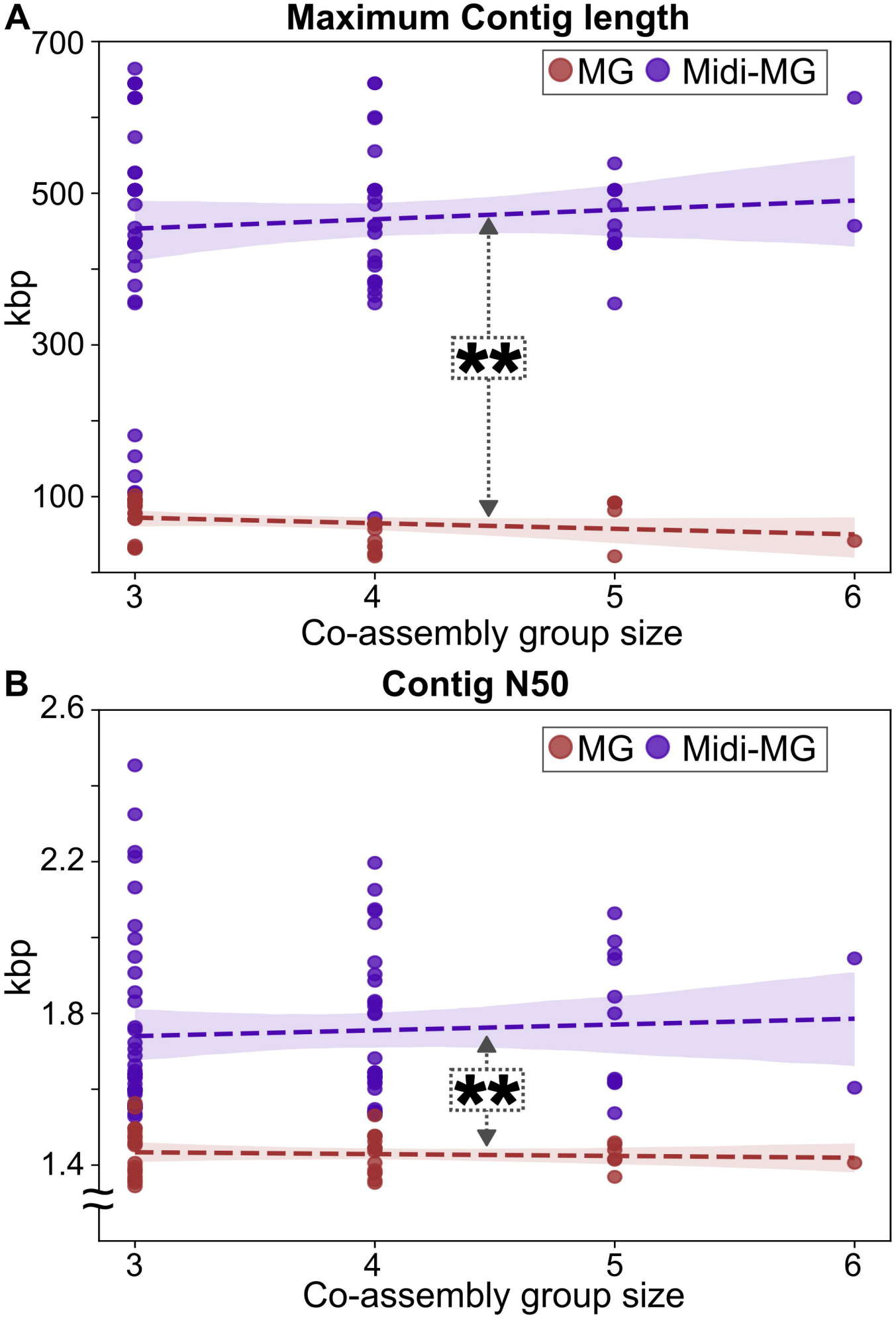
- Accurate coverage info
- Co-abundance binning
- Completeness higher than SCG and Mini-MG
- High complexity
- Binning is error prone

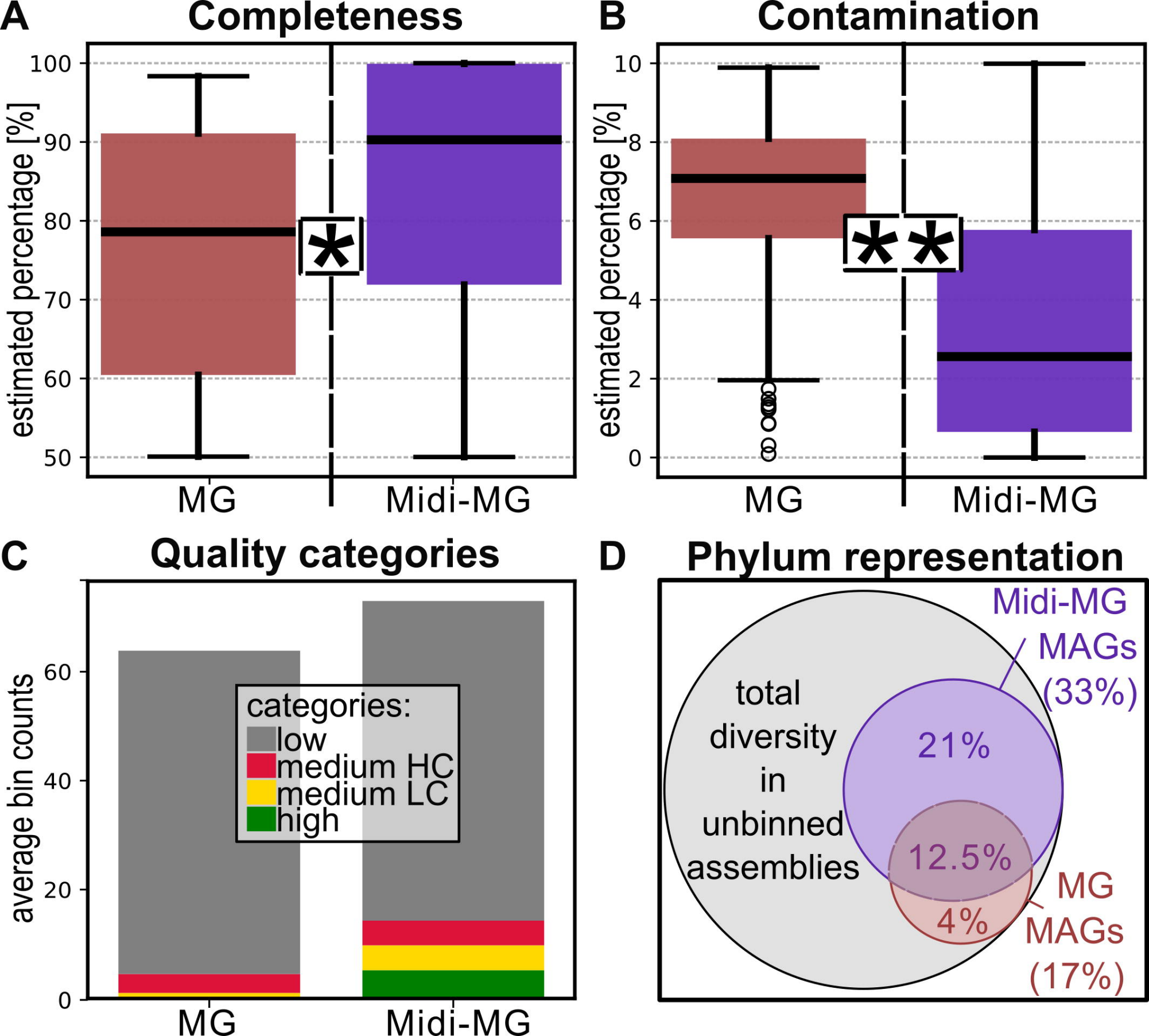
- Low complexity (single cells)
- No binning
- Completeness lower than MG and Mini-MG
- Contamination-prone WGA
- Coverage bias by WGA

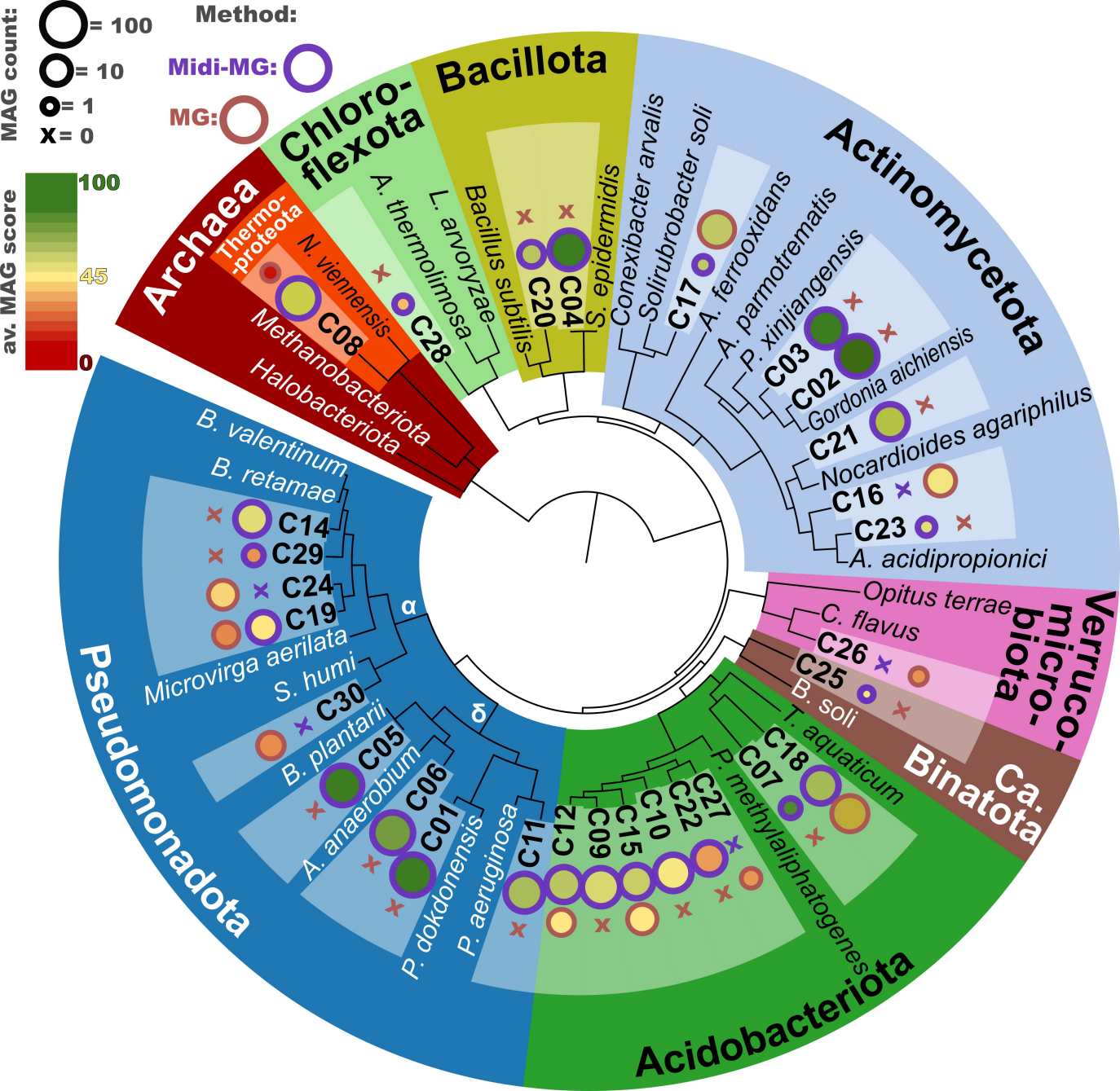
- Moderate complexity
- Completeness higher than SCG
- Completeness lower than MG
- Contamination-prone WGA
- Coverage bias by WGA
- Binning is error prone


sample
sorted fractions
entire community
A
cell sorting (FACS)
B
DNA extraction
C
sequencing
D
co-assembly
+
coverage info
functional & community profiles
MAGs


Pseudomonadota
Actinomycetota
Acidobacteriota
Planctomycetota
Gemmatimonadota
Bacteroidota
Chloroflexota
Verrucomicrobiota
Thermomicrobiota
Bacillota
Archaea
Eukaryota (Plant)
other
unclassified
Taxonomy:
UniFrac beta-diversity clustering cladogram
BNC
BC
SNC
MC
SC
Winter21/22
Big
Small
Autumn 20
BC
MC
SC
BNC
SNC
Summer21
sorted midi-metagenomic fractions
Spring20
Winter21/22
Winter22/23
Summer21
Autumn22
Autumn20
unsorted metagenomes
0
25
50
75
100
relative abundance [%]
Sampling timepoints:
Spring20
Autumn20
Summer21
Winter21/22
Autumn22
Winter22/23
Fractions:
= unsorted
= Small
= Big
= Small Non-Complex (SNC)
= Big Non-Complex (BNC)
= Small Complex (SC)
= Medium Complex (MC)
= Big Complex (BC)


A
Maximum Contig length
MG
Midi-MG
700
500
300
100
kbp
**
3
4
5
6
Co-assembly group size
B
Contig N50
MG
Midi-MG
2.6
2.2
1.8
1.4
kbp
**
3
4
5
6
Co-assembly group size


A
Completeness
estimated percentage [%]
100
90
80
70
60
50
*
MG
Midi-MG
B
Contamination
estimated percentage [%]
10
8
6
4
2
0
**
MG
Midi-MG
C
Quality categories
average bin counts
60
40
20
0
categories:
low
medium HC
medium LC
high
MG
Midi-MG
D
Phylum representation
total diversity in unbinned assemblies
Midi-MG MAGs (33%)
21%
12.5%
4%
MG MAGs (17%)


MAG count:
= 100
= 10
= 1
x = 0
Method:
Midi-MG:
MG:
av. MAG score
100
45
0
Archaea
Thermo-proteota
Halobacteriota
Methanobacteriota
N. viennensis
C08
Chloro-flexota
A. thermolimosa
C28
Bacillota
L. arvoryzae
Bacillus subtilis
C20
C04
S. epidermidis
Actinomycetota
Conexibacter arvalis
Solirubrobacter soli
C17
A. ferrooxidans
A. parmotrematis
P. xinjiangensis
C03
C02
Gordonia aichiensis
C21
Nocardioides agariphilus
C16
C23
A. acidipropionici
Verruco-micro-biota
Opitus terrae
C. flavus
C26
C25
B. soli
Ca. Binatota
T. aquaticum
C18
C07
P. methylaliphatogenes
Acidobacteriota
C27
C22
C10
C15
C09
C12
C11
P. aeruginosa
P. dokdonensis
C01
C06
A. anaerobium
C05
B. plantarii
C30
S. humi
Microvirga aerilata
C19
C24
C29
C14
B. retamae
B. valentinum
α
δ
Pseudomonadota