bioRxiv

"Midi-metagenomics": A novel approach for cultivation independent microbial genome reconstruction from environmental samples

John Vollmers*, Maximiano Correa Cassal and Anne-Kristin Kaster*
Institute for biological Interfaces 5, Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen
*****Correspondence should be addressed to john.vollmers@kit.edu and kaster@kit.edu

As the majority of environmental microbial organisms still evade cultivation attempts, genomic insights into many taxa are limited to cultivation-independent approaches. However, current methods of metagenomic binning 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 known to be prone to bias. Here we propose a novel approach for the purpose of metagenomic genome reconstructions, that utilizes the potential of fluorescence-activated cell sorting (FACS) for targeted enrichment and depletion of different cell types to create distinct cell fractions of sufficient size circumvent 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 of multiple distinct samples and has potential for the targeted enrichment and sequencing of uncultivated organism of interest.

Introduction

According to current estimates, less than 1% of environmental prokaryotes are culturable under laboratory conditions. The vast majority of microorganisms remain unavailable for direct analysis with classic microbiological methods and is thus commonly referred to as 'microbial dark matter'[1,2]. However, advances in cultivation-independent methodologies such as metagenomics and single-cell genomics nowadays enable thorough genome analyses of uncultured organisms[3,4] **(Figure 1).**

In metagenomics **(Figure 1A)**, the entire DNA of an environmental community is extracted, sequenced, and analysed[5]. Unfortunately, the assembly of individual discrete genomes from metagenomics data is sometimes not possible, especially for highly complex communities and organisms of low abundance. As a result, metagenome assembled genomes (MAGs) are highly susceptible to chimerism, meaning that they can contain contigs that originate from the genomes of different taxa[6,7].

Single-cell genomics (SCG) **(Figure 1B)** reduces this risk by targeting individual cells[2,8]. However, since a single bacterial cell contains only a few femtograms of DNA and the minimum requirement for high throughput sequencing is typically in the nanogram range, a whole genome amplification (WGA) is required. This is a severe disadvantage, as WGA usually yields extremely uneven read coverage, constituting a large bias that is particularly pronounced for genomes with high GC content and usually results in fragmented as well as incomplete single cell amplified genomes (SAGs)[2,9,10].

In order to minimize the individual 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, which targets small groups of usually 5-100 cells **(Figure 1C)**. These cells are then sequenced together

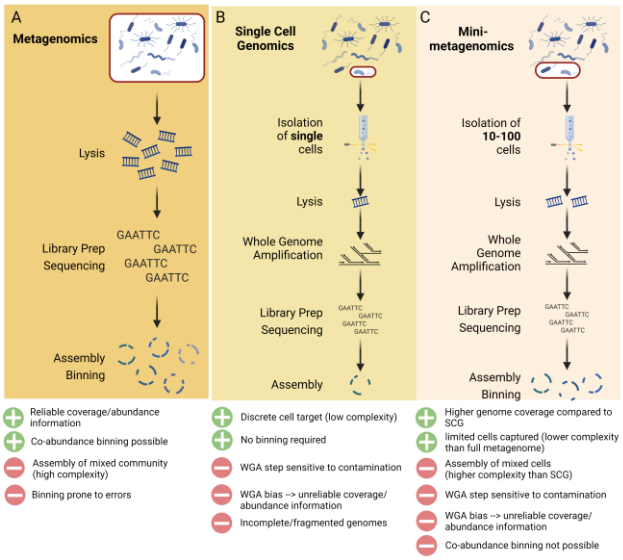

Figure 1: Current culture-independent methodologies. (A) Metagenomics utilizes the entire DNA of an environmental community for extraction, sequencing, and construction of metagenome assembled genomes (MAGs). Coverage information may be used to infer relative abundances and thus utilized for binning. However, assembly may be complex for highly diverse communities. **(B)** In **single-cell genomics (SCG)** individual cells are separated and sequenced, therefore assembly complexity is greatly reduced. However, since individual cells contain too little DNA for direct sequencing, whole genome amplification (WGA) steps need to be employed, introducing strong coverage bias. **(C) Mini-metagenomics** comprises of sorting pools of typically 5-100 cells and subjecting these to lysis, WGA and sequencing. By pooling multiple cells, random WGA bias is thought to be reduced while still maintaining a relatively low assembly complexity. However, the resulting assemblies nonetheless comprise multiple organisms while the remaining WGA bias prohibits abundance based binning efforts.

and subsequently treated as a simplified metagenome[9,11]. The DNA yield of such small cell groups is, however, still not sufficient to circumvent amplification, but is thought to efficiently reduce random bias. Furthermore, 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 be affected by systematic WGA bias that may be caused by e.g. variations in GC content[12]. 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[13]. Therefore contigs will likely have to be binned exclusively based on nucleotide signatures, which, however, can be unreliable, especially for short contigs of highly fragmented genomes[7].

Therefore, we here present a novel 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. We do so by utilizing the potential of fluorscence-activated cell sorting (FACS) 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 and Discussion

Basic principle of midi-metagenomics

In the here presented midi-metagenomic approach, the original sample population is divided into multiple fractions, in which different community members are selectively enriched or depleted **(Figure 2A**).

Selective fractionation is achieved *via* fluorescence-activated cell sorting (FACS). However, in contrast to standard single-cell and "mini-metagenomics", approaches which require an amplification step[9,11,14–16], several hundred thousand to million cells are sorted into the same fraction, enabling DNA yields in the multiple nanogram range sufficient for direct sequencing **(Supplementary Table S1)**. DNA is then extracted and sequenced **(Figure 2B&C)** from each fraction, as well as the original unfractionated sample, separately, resulting in multiple read datasets. Each of these datasets represents a different composition of the exact same original microbial community, in theory providing optimal conditions for subsequent co-assembly as well as co-abundance variation[17–20] based binning approaches **(Figure 2D**).

Furthermore, since each fraction simply represents a different view of the exact same community, this solves a common dilemma: Although co-assembly of multiple samples has been shown to increase genome recovery rates especially for low abundant species[21], it often also produces more fragmented assemblies and increases the risk of strain or species-level chimeras due to increased strain heterogeneity[22]. Such heterogeneities are often introduced by seasonal or locational variability between sample spots

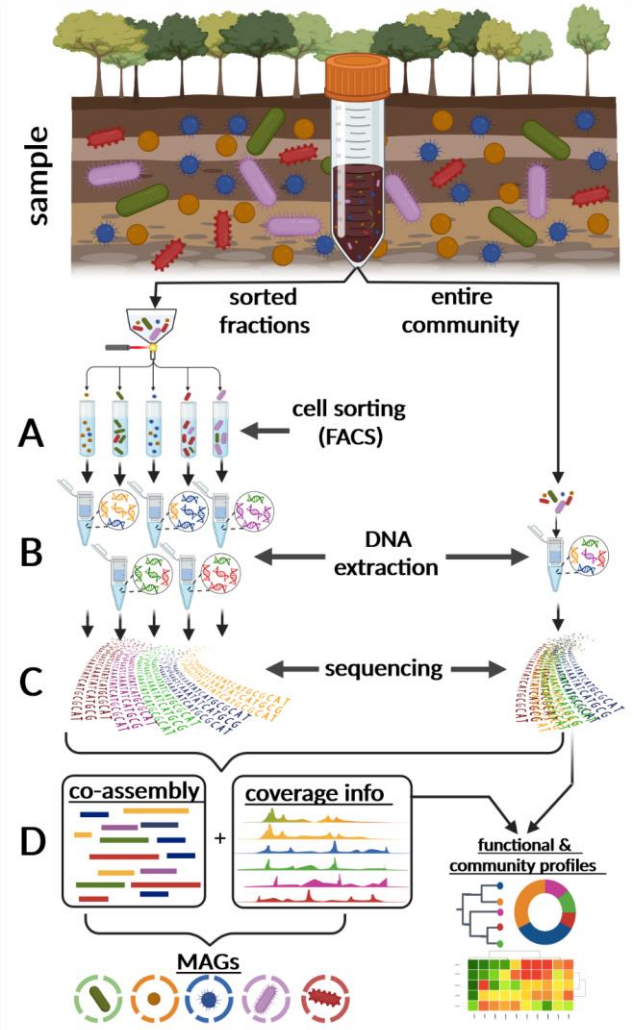

Figure 2: Midi-metagenomics workflow. (A) Part of the sample community is fractionated into distinct groups of several hundred thousand to millions of cells by Fluorescence-activated sorting (FACS). Different cell types are not separated with absolute stringency but differentially enriched **(B)** DNA is extracted separately from each fraction, as well as the original unsorted sample. **(C)** Extracted DNA is sequenced directly without applying 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

or sampling times. This variability is, however, also supposed to be exploited by co-abundance variation-based binning approaches. In the midi-metagenomics approach, the fact that all fractions originate from the same basic community prevents inter-sample strain variability, thereby maintaining optimal conditions for co-assembly.

Establishment and application of midi-metagenomics

Possible strategies for selectively fractionating a complex community into distinct subpopulations *via* FACS are manifold and can be based on phylogenetic, physiological or morphological properties of the target organisms[15], e.g., FISH staining using rRNA or mRNA probes[2,23,24], autofluorescence detection[25] or simply cell size and complexity[26]. In order to improve MAG reconstruction by

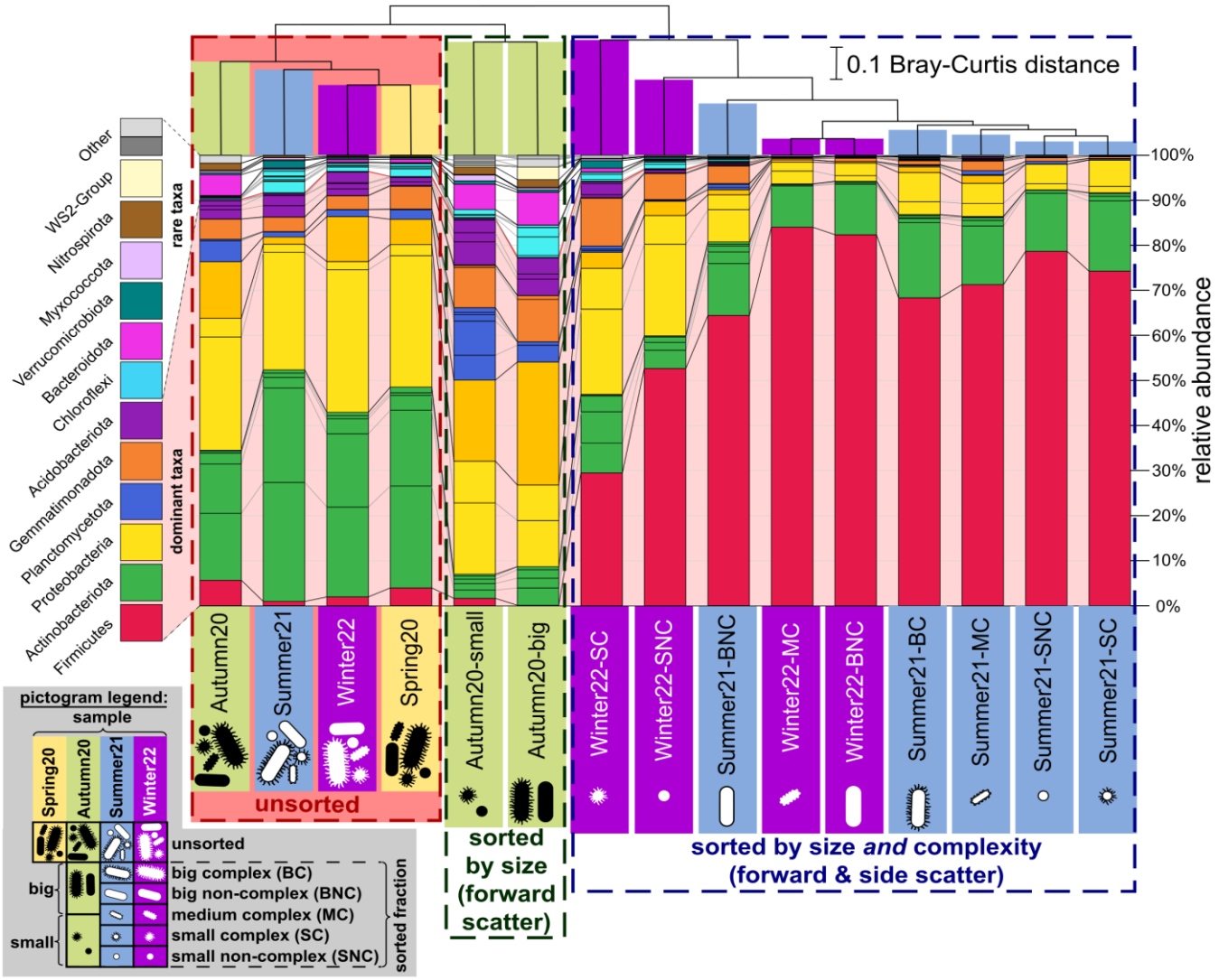

Figure 3: Differences in diversity between distinct unsorted samples as well as corresponding sorted fractions, determined based on amplicons of the 16S rRNA V3-region. Clustering reflects Bray-Curtis distances between samples and fractions (beta-diversity). Samples are indicated by background colouring, while fractions are indicated by pictograms, according to the legend at the lower left side. Stacked bar charts indicate the community composition of each sample and fraction, with different phyla being indicated by a distinct colour code as indicated on the left. Presence and relative abundance of distinct classes are indicated by black horizontal lines forming sub-sections within each stacked coloured bar. For the "Spring20" sample, only the unsorted complete community was analyzed. Only size fractions were generated for the first "Autumn20" sample, while size and complexity fractions were generated for the subsequent "summer21" as well as "winter22" samples. The "BC" fraction of "winter22" is not represented here, as the corresponding amplicon library yielded too few reads for thorough comparison. All amplicon libraries were based on the same extraction method described for sorted fractions (see **material & methods**) and the resulting read datasets were normalized to 45 000 read pairs each.

subsequent co-abundance variation-based binning approaches[20], fractionation does not need to be particularly stringent as long as a tendential enrichment or depletion can be achieved for at least some of the involved taxa. To illustrate this point, community fractionation in this proof-of-principle study was based on relatively simple gating strategies exploiting only those cell characteristics easily detectable *via* FACS: cell size and complexity as defined by forward- and side scatter gating, respectively **(Supplementary Figure S4)**. For the same reason, soil was chosen as a test subject, as it represents one the most complex and challenging microbial communities for metagenomic analyses[4,27].

The establishment of the involved DNA extraction protocol from sorted fractions was an iterative process. For the first sample ("spring20"), DNA extractions were attempted on cell pellets after centrifugation of sorted fractions. However, no DNA could be recovered from these pellets, therefore "spring20" is only included here as a non-fractionated sample for comparison of standard metagenomic co-assembly and binning approaches. Preliminary attempts indicated that after FACS and subsequent centrifugation of sorted cell fractions, DNA was present mostly in the supernatant and not the pellet[28] **(Supplementary Figure 1)**. The most likely explanation for this is cell damage due to stress caused by the sorting process and subsequent release of cellular DNA[28–31].

Accordingly, we modified the DNA extraction and included a DNA precipitation step directly from sorted cell suspensions, which resulted in successful DNA extractions

from sorted fractions beginning with the second sample, "autumn20". However, as 16S rRNA amplicon profiles based on universal bacterial primers (**material & methods**) showed a distinct lack of Actinobacteria and Firmicutes compared to the non-fractionated metagenomes despite using the same extraction method**(Figure 3),** we deduced that gram-positive cell walls may be more resilient to FACS stress and additional lysis steps may be required to capture cellular DNA from such organisms. Therefore, we introduced a simple bead beating step prior to DNA precipitation in the final extraction protocol that was applied for the subsequent "summer21" and "winter22" samples (see material and methods). This modification lead to a strong representation of Actinobacteria and even an apparent overrepresentation of Firmicutes in the amplicon profiles of the resulting sorted fractions **(Figure 3)**. The latter observation may be predominantly attributable to the fact that standard metagenomes even also capture cell free DNA, while FACS can only capture organisms that were at least intact enough to maintain cellular shape: Since Firmicutes were exclusively represented by members of the class Bacilli, which are known as proliferate spore formers, the higher stability of bacillus spores compared to vegetative cells readily explains their high representation in the sorted fractions. While this effect may be avoidable by employing live/dead staining during FACS sorting, it can actually be used to advantage for co-abundance variation-based binning, especially since the exact degree of overrepresentation varied between fractions.

With the final adapted extraction method, the DNA yield obtained from fractions of up to 5 million cells ranged between 5-30 ng (**Supplementary Table S1, Supplementary Figure S5**). We note that, since sequencing library preparation is nowadays already possible with less than a nanogram of input DNA, sorting efforts may be significantly reduced to fractions of down to 100 000 cells.

Community Fractionation

We also analyzed how distinct the sorted fractions are compared to each other, as well as to the corresponding unsorted samples, based on 16S rRNA amplicons. Bray-Curtis distance values 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 years and seasons **(Figure 3)**, despite employing the exact same DNA extraction protocoll for all amplicon libraries. 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. Interestingly, the simple size-based fractions of autumn20 show a higher similarity to unsorted samples than the more detailed size and complexity-based fractions of "summer21" and "winter22". This illustrates that the distinctness of each sub-population is linked to the degree/detail to which the community was fractionated and shows promise that this can be fine-tuned to individual requirements.

It needs to be noted that the sorted fractions clearly cluster closer together to other sorted fractions than to their respective unsorted samples, which also indicates a general influence of the FACS process. However, by always including an unsorted sample in the analyses, an unbiased view of the original community is maintained and any systematic influences from the sorting process can simply be exploited for binning purposes.

The difference of beta-diversity distances between sorted fractions varied strongly, with the largest Bray-Curtis dissimilarity values being observed between the "Small Complex" (SC) and respective "Big Non-Complex" (BNC) fractions. Therefore, in order to minimize effort, when fractionating based on simple size and complexity

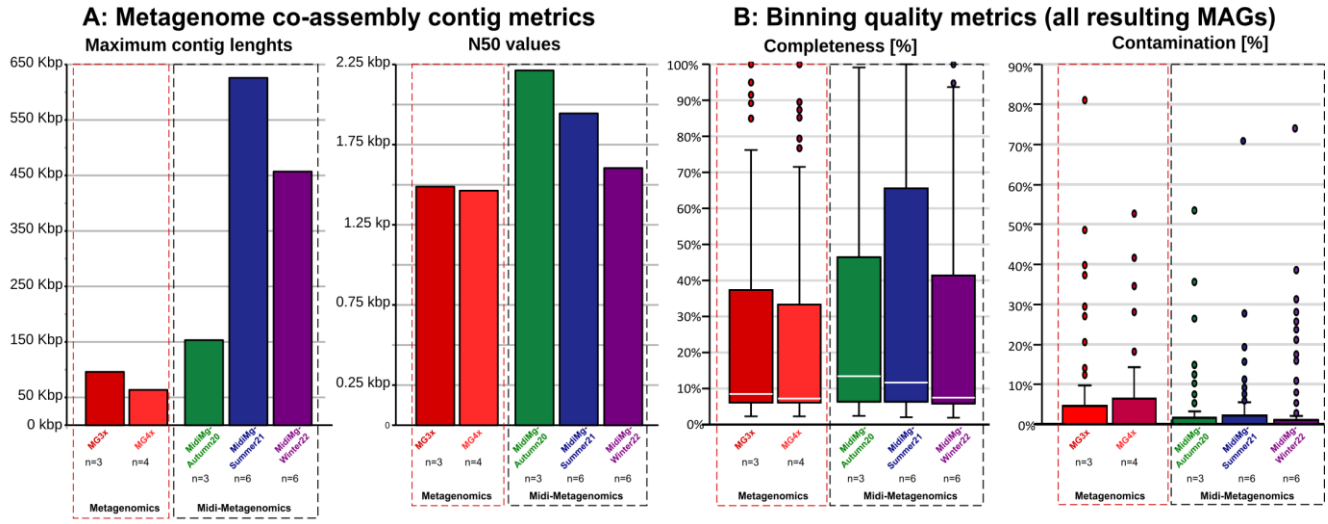

Figure 4: Comparison of overall assembly and average bin quality netrics for the different approaches and samples. (A) General contig size metrics of the (midi-)metagenome co-assemblies *before* binning. **(B)** Distribution of basic MAG quality metrics for all MAGs obtained from the co-assemblies *after* binning and clean-up. MAG quality metrics were derived via checkm2[49]. The number of samples and fractions incorporated into each co-assembly are as follows: "MG3x" = 3 samples with 1 fraction each; "MG4x" = 4 samples with 1 fraction each; "midiMG-autumn20" = 1 sample, 3 fractions; "MidiMG-summer21" & "-winter22" = 1 sample, 6 fractions each. The exact composition of samples and fractions incorporated into each co-assembly is listed in **Table 2** and **Supplementary Figure S2**.

characteristics, beta diversity can already be maximized by sorting just these two fractions.

Several taxa appear to show strong fluctuations of cell size and complexity between sampling timepoints, in particular members of the phyla Firmicutes, Proteobacteria, Gemmatimonadota and Planctomycetota (**Figure 3; Supplementary Tables S2-S4**). Furthermore, in the majority of these cases, a strongly pronounced difference between the "winter22" and "summer21" samples is noticeable. The clearest example is given by members of the class Bacilli, which are enriched in the "Big" and "Non-Complex" fractions in the "summer21" sample but show the exact opposite trend in the "winter22" sample, with enrichment in the "Small" and "Complex" fractions (**Figure 3; Supplementary Tables S2-S4**).

Interestingly, this exact relationship is also displayed on individual OTU level e.g., for multiple members of the genus *Bacillus* (**Supplementary Tables S2 and S5-S8**), showing that this is not simply caused by succession of different related species, but that actual individual strains vary in size and cell shape between samples of different timepoints. This is not necessarily surprising, especially for members of the genus *Bacillus*, which are known to form endospores in reaction to different environmental conditions[32,33] or Planctomycetes which are known to display complex cell cycles[24]. Furthermore, fluctuation of cell size in dependence of season and nutrient availability has already been reported for multiple taxa in various environments[34–37]. This effect clearly illustrates the potential of the midi-metagenomic approach to capture and exploit the fluctuation of specific cell characteristics over time or across environmental conditions.

Assembly and binning performance

We compared co-assemblies of metagenomic and midi-metagenomic approaches, using always the same total sequencing depth of 15 Gbp (averaging at 70 mio read pairs per co-assembly) equally distributed across the respective combined samples and fractions (table 2). Based on N50 and maximum contig length metrics, the standard metagenome co-assemblies, consisting of different samples from different years and seasons, were generally more fragmented than the midi-metagenomic co-assemblies consisting of multiple fractions originating from the same respective samples **(Figure 4)**. Co-assembly of multiple datasets poses a common dilemma in metagenomics: Although co-assembly of multiple samples has been shown to increase genome recovery rates especially for low abundant species[21], it often also produces more fragmented assemblies and increases the risk of strain or species-level chimeras due to increased strain heterogeneity which can be introduced by seasonal or locational variability between samples directly affecting the complexity of the assembly-graph [22].

Accordingly, the multi-sample co-assembly of four soil samples (STD4) was slightly more fragmented that with only three samples (STD3) with slightly lower N50 and maximum contig length values (**Figure 4A**). In all cases, midi-metagenomic co-assemblies yielded larger, less fragmented contigs than the multi-sample co-assemblies, regardless of whether only three or even six read datasets were co-assembled. This indicates that the midi-metagenomic approach of only co-assembling fractions of the same sample allows the maximization of read coverage while avoiding the typical increase of complexity that is introduced by combining different samples, e.g. though inter-

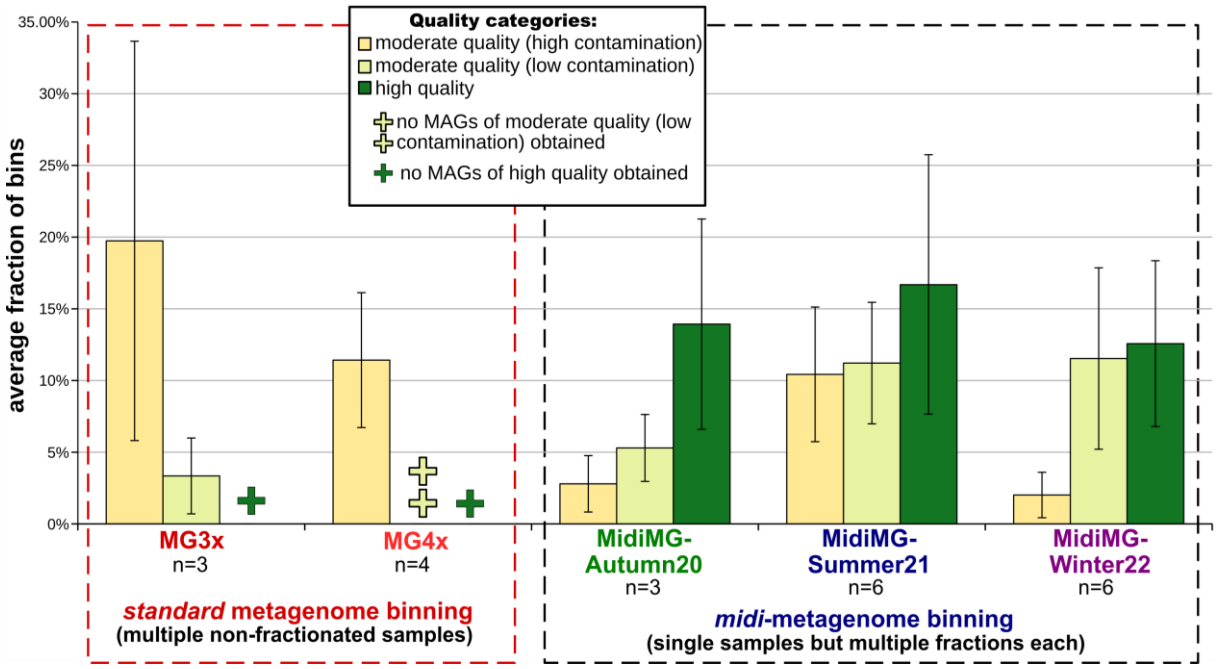

Figure 5: Relative proportions of moderate and high-quality MAGs obtained by standard and midi-metagenomic approaches. The primarily size based MIMAGS category of "moderate quality" was further divided into subgroups with "high" (5-10%) and "low" (0-5%) contamination estimates, in order to place higher emphasis on this critical metric. The total number of samples and fractions *n* combined in each assembly are indicated below the respective assembly designation while the exact composition is listed in **Table 2** and **Supplementary Figure S2**.

sample strain variations. Therefore, by distributing sequencing efforts over multiple fractions, both coverage variation and sequencing depth can be maximized at the same time in a cost-efficient manner.

Expectedly, improved assembly metrics also affect the quality of the produced MAGs: Midi-metagenomic MAGs show tendentially higher completeness values, with upper quartiles ranging between 40-65% and at the same time lower contamination estimates, with upper quartiles ranging between 1 and 2%, compared to their counterparts from standard multi-sample metagenome co-assemblies where upper quartiles of completeness and contamination values were 5% and 32-37%,respectively (**Figure 4B**). These improvements become most obvious when eliminating MAGs of "low quality" and contamination values above 10%, which are of limited scientific interest[7,38], and instead

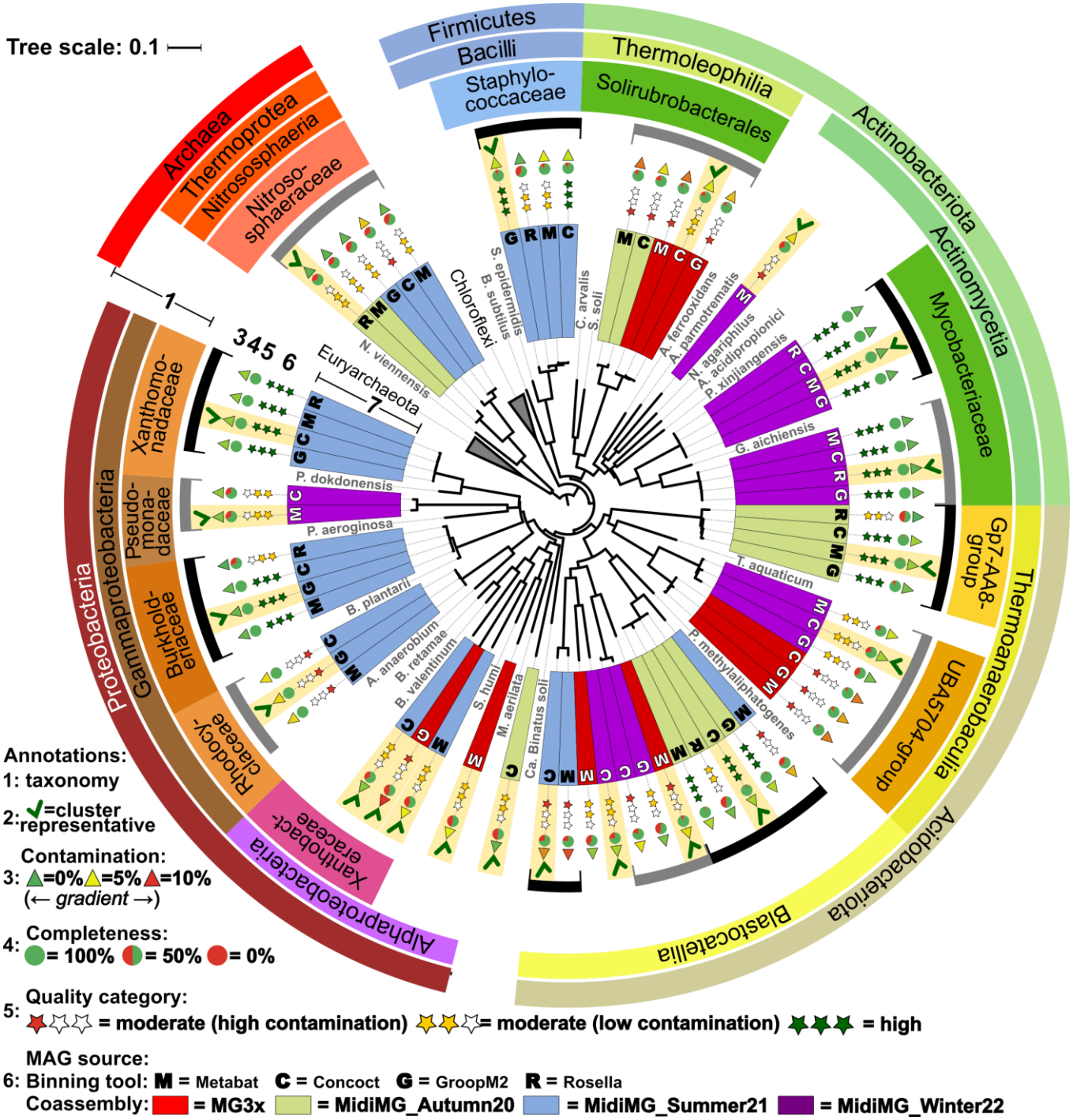

Figure 6: Gene content clustering of MAGs fulfilling at least the "moderate" quality category of MIMAG standards as well as selected reference genomes. Coloured boxes indicate the assembly that produced the respective MAGs, while the respective binning tool is indicated by bold letters, according to the legend on the lower left. Adjusted MIMAG[38] quality category, completeness, contamination as well as general taxonomic affiliation are indicated by outer annotation rings according to the legend. Clusters of MAGs that likely represent the same species according to dRep[22] analyses are indicated by black and grey brackets. The respective representative MAG showing the highest quality per potential species is marked by green ticks and highlighted in yellow. In order to place more emphasis on purity than completeness, the "moderate" quality category of MIMAG standards was further subdivided into "low contamination" and "high contamination", representing 0-5% and 5-10% contamination estimates, respectively.

focusing on the fraction of moderate and high quality genomes obtained from the different approaches (**Figure 5**): Using the standard multi-sample approach, not a single MAG fulfilling MIMAGs "high quality" criteria [38] could be reconstructed from the soil metagenome co-assemblies with any of the tested binning tools[13,18,38–40]. On the other side, several "high quality" bins could be produced by most binning tools when using the midi-metagenomics approach.

It is important to note that for quality categories below "high quality", MIMAG standards are unfortunately not adequate, as they do not differentiate between different degrees of contamination. Instead, the current standards group all MAGS, even up to 10% contamination, into the same "quality groups", which are primarily based on completeness estimations[38]. Since this is counterproductive to efforts that aim to minimize gradual reference database contaminations, it is crucial to place high emphasis on contamination besides completeness[7]. Consequently, we sub-categorized all MAGs fulfilling MIMAG criteria of "moderate quality" into those with low contamination (<5%) and those with high contamination (5-10%). With these added criteria it becomes clear that the midi-metagenomic approach produces far larger numbers of representative MAGs, with 11-16 "low contamination" MAGS of moderate quality or higher, compared to only 0-3 produced by the standard multi-sample metagenome approaches **(Figures 4 - 6,Supplementary Table S9)**.

A total of 69 MAGs were obtained from all sampling, assembly and binning approaches combined, that could be classified as "moderate" to "high" quality. Several of these form clusters of potentially redundant MAGs likely representing the same species, as determined by dRep[22] and confirmed *via* gene-content-based clustering, **(Figure 6)**. Interestingly, in the vast majority of these clusters of potentially redundant MAGs, the most representative MAG determined by dRep was provided by a midi-metagenomic rather than a standard metagenomic approach, reflecting the generally higher quality results of midi-metagenomics. Furthermore, whenever gene-content analyses indicated sub-clades within such clusters of redundant MAGs, these represented different samples, illustrating a high potential of midi-metagenomics to capture and resolve inter-sample strain heterologies. Also noteworthy is the observation that with up to two domains and five different classes the moderate to high quality MAGs of the midi-metagenomic approaches cover a much higher phylogenetic range than the standard approach, which represents only three bacterial classes. This is predominantly due to the fact that most of the MAGs produced by the multi-sample approaches are of "low quality" or even disqualify completely due to contamination values beyond the acceptable range defined by MIMAGs[38] **(Supplementary Figure S3).**

Conclusion

We could here show that the midi-metagenomic approach of combining sequence information from multiple fractions of the same sample produces higher quality assemblies and MAGs compared to the classic metagenomic method of combining sequence information of multiple distinct samples.

In order to achieve these advantages, sorting criteria do not even require particularly high stringency as long as simply partial enrichment or depletion can be achieved, allowing setups to be kept minimal and simple. In fact, just the simple act of FACS itself already represents a general depletion of large multi-cell aggregates, extracellular DNA as well as potential cell types that may be more susceptible to FACS stress[28]. Of course, this also means that in order to include an accurate representation of the actual natural community, unsorted shotgun metagenomes libraries always need to be created and included as well. However, since multiple fractions can be sorted from a single sample, depending on the sample and the exact research goal, i.e. the importance of capturing seasonal or spatial variation, the overall effort is not necessarily higher than standard multi-sample approaches. Furthermore, midi-metagenomics may serve to boost binning efforts in cases where the variability between samples may turn out not be sufficient for co-abundance based binning, especially for sampling locations that are hard or expensive to access for additional sampling trips, i.e. deep-sea vents.

In this proof-of-principle study we achieved substantial improvements in the quality of MAGs obtained from highly complex soil communities, just by generating simple size and complexity-based fractions. The maximum potential of this approach however will likely be realized when other cell properties can also be exploited to produce more distinct community fractions. A most simple improvement could e.g., be the addition of live-dead staining, which would distinguish spores and damaged or dead cells from viable ones. Furthermore, previously established 16S rRNA FISH labelling could help to more stringently target specific taxa for enrichment or depletion[23]. On functional level, FISH-labelling could also be employed to simply separate cells by rate of metabolic activity based on ribosome-content using broad-range bacterial/archaeal rRNA probes. Depending on the research question it should even be possible to enrich cells expressing certain genes of interest, by using specific mRNA targeting probes instead of rRNA probes. Last but not least, autofluorescence spectra caused by diverse membrane proteins in different organisms[25] may be exploited to enrich or deplete cells with specific functional properties.

In this regard, midi-metagenomics has the added potential to specially target and analyze specific organisms, functions or metabolic traits of interest within complex communities, regardless of actual abundance in the sample. The exact sorting criteria do not even need to be known before or during sampling as a glycerol stock of frozen sample can be revisited for sorting even after preliminary whole-community metagenome analyses.

The most significant advantage of the midi-metagenomics approach, however, is that it allows the maximization of sequence data for assembly as well as co-abundance variation-based binning purposes, while simultaneously avoiding the complications typically introduced by inter-sample strain heterologies. This means that read depth may be distributed across multiple fractions without negatively affecting assembly quality, therefore potentially reducing the required sequencing cost. With the omission of highly expensive whole genome amplification

(WGA) techniques[2], this method also provides a cost-efficient alternative to single cell genomics and mini-metagenomics approaches.

The result are more diverse MAGs that better represent the respective organisms of interest with substantially less contamination compared to traditional metagenome binning approaches. This is of particular significance as the minimization of MAG contamination desperately needs to be prioritized, considering recent complaints of increasing reference database contaminations caused by insufficiently screened MAGs and SAGs[7,39,40].

References

1. Bodor, A. *et al.* Challenges of unculturable bacteria: environmental perspectives. *Rev Environ Sci Biotechnol* **19**, 1–22 (2020).
2. Kaster, A. K. & Sobol, M. S. Microbial single-cell omics: the crux of the matter. *Appl Microbiol Biotechnol* **104**, 8209–8220 (2020).
3. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).
4. Vollmers, J., Wiegand, S. & Kaster, A. K. Comparing and evaluating metagenome assembly tools from a microbiologist's perspective - Not only size matters! *PLoS One* **12**, 1–31 (2017).
5. Schmeisser, C., Steele, H. & Streit, W. R. Metagenomics, biotechnology with non-culturable microbes. *Appl Microbiol Biotechnol* **75**, 955–962 (2007).
6. Orakov, A. *et al.* GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome Biol* **22**, 1–19 (2021).
7. Vollmers, J., Wiegand, S., Lenk, F. & Kaster, A.-K. How clear is our current view on microbial dark matter? (Re-)assessing public MAG & SAG datasets with MDMcleaner. *Nucleic Acids Res* **50**, e76 (2022).
8. Xu, Y. & Zhao, F. Single-cell metagenomics: challenges and applications. *Protein Cell* **9**, 501–510 (2018).
9. Alteio, L. v. *et al.* Complementary metagenomic approaches improve reconstruction of microbial diversity in a forest soil. *mSystems* **5**, 1–18 (2020).
10. Lasken, R. S. & Stockwell, T. B. Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnol* **7**, 1–11 (2007).
11. Yu, F. B. *et al.* Microfluidic-based mini-metagenomics enables discovery of novel microbial lineages from complex environmental samples. *Elife* **6**, 1–20 (2017).
12. Marine, R. *et al.* Caught in the middle with multiple displacement amplification: the myth of pooling for avoiding multiple displacement amplification bias in a metagenome. *Microbiome* **2**, 3 (2014).
13. Alneberg, J. *et al.* CONCOCT: Clustering cONtigs on COverage and Composition. *ArXiv* **1312**, 1–28 (2013).
14. Tracy, B. P., Gaida, S. M. & Papoutsakis, E. T. Flow cytometry for bacteria: enabling metabolic engineering, synthetic biology and the elucidation of complex phenotypes. *Curr Opin Biotechnol* **21**, 85–99 (2010).
15. Woyke, T., Doud, D. F. R. & Schulz, F. The trajectory of microbial single-cell sequencing. *Nat Methods* **14**, 1045–1054 (2017).
16. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).
17. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* **31**, 533–8 (2013).
18. Imelfort, M. *et al.* GroopM : an automated tool for the recovery of population genomes from related metagenomes. *PeerJ* **2**, e603 (2014).
19. Lin, H.-H. & Liao, Y.-C. Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. *Sci Rep* **6**, 1–8 (2016).
20. Nielsen, H. B. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature Biotechnology 2014 32:8* **32**, 822–828 (2014).
21. Hofmeyr, S. *et al.* Terabase-scale metagenome coassembly with MetaHipMer. *Scientific Reports 2020 10:1* **10**, 1–11 (2020).
22. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J* **11**, 2864–2868 (2017).
23. Dam, H. T., Vollmers, J., Sobol, M. S., Cabezas, A. & Kaster, A.-K. Targeted cell sorting combined with single cell genomics reveals microbial dark matter overlooked by metagenomics. *Front Microbiol* **11**, 1377 (2020).
24. Pratscher, J., Vollmers, J., Wiegand, S., Dumont, M. G. & Kaster, A. K. Unravelling the identity, metabolic potential and global biogeography of the atmospheric methane-oxidizing upland soil cluster α. *Environ Microbiol* **20**, 1016–1029 (2018).
25. Kang, S. M., de Josselin de Jong, E., Higham, S. M., Hope, C. K. & Kim, B. il. Fluorescence fingerprints of oral bacteria. *J Biophotonics* **13**, (2020).
26. Galbusera, L., Bellement-Theroue, G., Urchueguia, A., Julou, T. & van Nimwegen, E. Using fluorescence flow cytometry data for single-cell gene expression analysis in bacteria. *PLoS One* **15**, (2020).
27. Jansson, J. K. & Hofmockel, K. S. The soil microbiome — from metagenomics to metaphenomics. *Curr Opin Microbiol* **43**, 162–168 (2018).
28. Wiegand, S., Dam, H. T., Riba, J., Vollmers, J. & Kaster, A.-K. Printing microbial dark matter: using single cell dispensing and genomics to investigate the Patescibacteria/Candidate Phyla Radiation. *Front Microbiol* **12**, (2021).
29. Binek, A. *et al.* Flow cytometry has a significant impact on the cellular metabolome. *J Proteome Res* **18**, 169–181 (2018).
30. Blainey, P. C. The future is now: single-cell genomics of bacteria and archaea. *FEMS Microbiol Rev* **37**, 407–427 (2013).
31. Mollet, M., Godoy-Silva, R., Berdugo, C. & Chalmers, J. J. Computer simulations of the energy dissipation rate in a fluorescence-activated cell sorter: Implications to cells. *Biotechnol Bioeng* **100**, 260–272 (2008).
32. Baril, E. *et al.* Sporulation boundaries and spore formation kinetics of *Bacillus* spp. as a function of temperature, pH and a_w. *Food Microbiol* **32**, 79–86 (2012).
33. Logan, N. A. & de Vos, P. *Bacillus*. in *Bergey's Manual of Systematics of Archaea and Bacteria* (ed. Whitman, W. B.) vol. 3 1–163 (John Wiley & Sons, Ltd, 2015).
34. Chien, A. C., Hill, N. S. & Levin, P. A. Cell size control in bacteria. *Current Biology* **22**, R340–R349 (2012).
35. Huete-Stauffer, T. M., Arandia-Gorostidi, N., Alonso-Sáez, L. & Morán, X. A. G. Experimental warming decreases the average size and nucleic acid content of marine bacterial communities. *Front Microbiol* **7**, 1–13 (2016).
36. Yao, Z., Davis, R. M., Kishony, R., Kahne, D. & Ruiz, N. Regulation of cell size in response to nutrient availability by fatty acid biosynthesis in *Escherichia coli*. *Proc Natl Acad Sci U S A* **109**, (2012).
37. Zohary, T., Fishbein, T., Shlichter, M. & Naselli-Flores, L. Larger cell or colony size in winter, smaller in summer–a pattern shared by many species of Lake Kinneret phytoplankton. *Inland Waters* **7**, 200–209 (2017).
38. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* **35**, 725–731 (2017).
39. Arkhipova, I. R. Metagenome Proteins and Database Contamination. *mSphere* **5**, e00854-20 (2020).
40. Breitwieser, F. P., Pertea, M., Zimin, A. v. & Salzberg, S. L. Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome Res* **29**, 954–960 (2019).
41. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
42. Bushnell, B. BBtools software package (36.84). *Sourceforge* https://sourceforge.net/projects/bbmap/ (2014).
43. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* **17**, 10–12 (2011).
44. Magoč, T. & Salzberg, S. L. FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).
45. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: A versatile open source tool for metagenomics. *PeerJ* **2016**, (2016).
46. Pruesse, E., Peplies, J. & Glöckner, F. O. SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**, 1823–1829 (2012).
47. Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
48. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).
49. Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *bioRxiv* 2022.07.11.499243 (2022) doi:10.1101/2022.07.11.499243.
50. Chaumeil, P. A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: A toolkit to classify genomes with the genome taxonomy database. *Bioinformatics* **36**, 1925–1927 (2020).
51. Howat, A. M. *et al.* Comparative genomics and mutational analysis reveals a novel XoxF-utilizing methylotroph in the Roseobacter group isolated from the marine environment. *Front Microbiol* **9**, 1–12 (2018).

Acknowledgements

The Authors acknowledge support by the state of Baden-Württemberg through bwHPC. The authors furthermore want to acknowledge the support by dr. Florian Lenk in proofreading this manuscript and providing helpful suggestions.

Author contributions

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

Competing interest statement

The authors declare that they have no competing interests

Materials and Methods

Microbial Samples

To evaluate midi-metagenomics performance compared to metagenomics, soil samples were collected in 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 Fluorescence-Activated Cell Sorting (FACS) 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 FACS 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 machine (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. Such 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" and "winter22", cells were sorted into five different groups according to their size and complexity, which are roughly proportional to the Forward Scatter Signal (FSC) and Side Scatter Signal (SSC), measured respectively by the cytometry lasers of the FACS machine (**Supplementary Table S1 & Supplementary Figure S1**). For sample "autumn20" only two groups were sorted, according to size measured by differences in FSC (**Supplementary Table S1**). 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 the unsorted soil samples for Metagenomics, DNA was extracted with 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)

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

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.

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
winter22		15.02.2022	unsorted, BC, MC, SC, BNC & SNC

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[41–43]: . Overlapping read pairs were identified and merged using FLASH v.1.2.11[44]. For amplicon datasets, reads were subsamples to 45 000 reads per dataset and operational taxonomic units (OTUs) were determined by read clustering using the unoise approach of VSEARCH v.2.21.1[45] and subsequently taxonomically classified using SINA v1.7.2[46]. Shotgun datasets were combined into co-assembly groups representing three, four or six datasets of either multiple samples, or multiple fractions of the same sample (**Table 2**). Each dataset was randomly subsampled down to 2.5 Gbp or 5 Gbp, depending on the size of the respective co-assembly group, ranging from three to six datasets, in order to achieve an equal amount of 15 Gbp sequencing data in total for each co-assembly. Co-assemblies were then performed using MegaHit v1.2.9 [47].

Table 2: Overview of assemblies. Fraction abbreviations: BC = "Big Complex", MC = "Medium Complex", SC = "Small Complex", BNC = "Big Non-Complex", SNC = "Small Non-Complex"

Assembly	Total input	Samples	Fractions
MG3x	15 Gbp	3 (autumn20, summer21, winter22)	3 (only unsorted)
MG4x	15 Gbp	4 (spring20, autumn20, summer21, winter22)	4 (only unsorted)
MidiMG-autumn20	15 Gbp	1 (autumn20)	3 (unsorted, Big & Small)
MidiMG-summer21	15 Gbp	1 (summer21)	6 (unsorted, BC, MC, SC, BNC & SNC)
MidiMG-winter22	15 Gbp	1 (winter22)	6 (unsorted, BC, MC, SC, BNC & SNC)

MAG reconstruction and analyses

For each co-assembly, four different binning tools were used in parallel: Metabat2 v.2.15, Concoct, GroopM2 v.2.0.0 and Rosella v.0.4.1[13,18,48]. Resulting bins were pre-assessed and filtered using MDMcleaner. Quality categories were then determined based on re-assessments using checkm2[49]. Taxonomic classifications were based on GTDB-TK v2.1.1[50]

dRep v.3.4.0 [22] 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[51].

Data availability

All sequencing data, co-assemblies as well as all MAGs of "high quality" or "moderate quality" with contamination estimates below 5% are available at NCBI under bioproject PRJNA900514. MAGs of low quality or with high contamination rates were not submitted to any dedicated sequence database in order to prevent gradual database corruption, but may be accessed via zenodo under the DOI: 10.5281/zenodo.7547690