Draft manuscript

UniteM: recovery of metagenome-assembled genomes using an ensemble of binning methods

Donovan H. Parks[1]*, Steve J. Robbins[1], Rochelle M. Soo[1], Paul N. Evans[1], Andy O. Leu[1], Gene W. Tyson[1]

[1]Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, St Lucia, QLD 4072, Australia

*Correspondence should be addressed to Donovan Parks (donovan.parks@gmail.com)

Recovering microbial genomes directly from environmental samples provides an opportunity to study organisms that are recalcitrant to cultivation and to further our understanding of the role of microorganisms in biogeochemical and industrial processes. Current methods for obtaining metagenome-assembled genomes often produce inconsistent results due to differences in their underlying models. Here we exploit the varying strengths of these models to obtain additional and more accurate genomes than provided by any individual binning method. We demonstrate the advantages of three different ensemble binning strategies on a simulated microbial community, a deep aquifer metagenome, and a diverse set of 1,550 public metagenomes. A flexible software package, UniteM, is provided that allows for easy binning across several popular binning methods and implements the three proposed ensemble binning strategies. We show that UniteM can provide additional insights into microbial communities through the recovery of additional near-complete genomes and by establishing the stability of metagenome-assembled genomes across different binning methods.

Introduction

Microbial communities play a central role in many industrial processes and are critical for proper ecosystem function. Metagenome assembled-genomes (MAGs) allow metabolic pathways to be reconstructed for *in situ* microbial populations and provide an opportunity to study organisms

that are recalcitrant to cultivation (Tyson et al. 2004; Brown et al., 2015; Evans et al., 2015). Current sequencing and assembly techniques allow MAGs of even low abundance microbial populations to be recovered from high diversity environments (Yeoh et al., 2016). However, alternative methods for obtaining MAGs may recover genomes from different populations and MAGs of the same population can vary substantially in quality (Sczyrba et al., 2017; Graham et al., 2017).

MAGs are obtained by clustering or "binning" together assembled contigs with similar sequence composition, depth of coverage across one or more related samples, and taxonomic affiliations (Albertsen et al., 2013; Sangwan et al., 2016). A number of approaches have been proposed for obtaining MAGs that vary in their underlying models for grouping assembled contigs into population-specific genomes (Strous et al., 2012; Wrighton et al., 2012; Imelfort et al., 2014; Nielsen et al., 2014; Kang et al., 2015; Wu et al., 2016). Despite substantial progress, the task of recovering genomes from metagenomic data remains challenging and MAGs are typically incomplete and may contain DNA from multiple populations. This has led to the development of methods for estimating the completeness and contamination of MAGs in order to help guide biological inferences being made from these genomes (Parks et al., 2015; Eren et al., 2015; Tennessen et al., 2016).

Recognition that different binning methods may recover complementary sets of MAGs has led to the development of methods for determining a non-redundant set of MAGs across multiple binning methods (Sieber et al., 2017; Song and Thomas 2017). Using an ensemble of clustering methods to obtain an improved or consensus clustering is a well-studied area in computer science (Strehl and Ghosh 2002; Fred and Jain, 2005). However, combining different binning methods deviates from classical ensemble clustering as there is a clear definition of what constitutes a correct MAG and methods exist for robustly estimating the quality of individual MAGs (Parks et al., 2015). This estimate of MAG quality is directly exploited by DAS Tool which iteratively selects the highest quality MAG across a set of binning methods in order to try and maximize the number of near-complete MAGs obtained from a metagenome (Sieber et al., 2017). Bin_refiner takes a more conservative approach that aims to minimize contamination by partitioning MAGs into refined bins consisting of contigs that are clustered together under all binning methods (Song and Thomas 2017).

Here we introduce a software suite, UniteM, which combines the output of different binning methods using three distinct strategies. The 'greedy' strategy is similar to DAS Tool, but uses a more robust estimate of genome quality when iteratively determining the highest quality MAG. The 'consensus' strategy identifies MAGs representing the same microbial population across multiple binning methods and takes a consensus vote of the contigs comprising these MAGs in order to produce a refined MAG of the population. The 'unanimous' strategy aims to minimize contamination by producing refined bins that represent a strict consensus across MAGs from different binning methods identified as representing the same microbial population. This strategy is similar in intent to Bin_refiner though only requires consensus across MAGs identified as representing the same population, while Bin_refiner requires consensus across all MAGs regardless as to whether or not they represent that same microbial population. We demonstrate the advantages of these different strategies on a simulated microbial community, through a detail examination of a deep aquifer metagenome, and over a diverse set of 1,550 previously analyzed public metagenomes.

Results

Complementary MAGs are Recovered by Different Binning Methods

We begin by demonstrating that existing binning methods can produce complementary MAGs by applying BinSanity v1 (Graham et al., 2007), GroopM v2 (Imelfort et al., 2014), MaxBin v2 (Wu et al., 2015), and MetaBAT v1 and v2 (Kang et al., 2015) under different parameter settings to the CX10 deep aquifer metagenome (Evans et al., 2015). Each of these binning methods was applied to the CX10 community with differential coverage information provided by 10 other deep aquifer metagenomes sampled from the same geographic location as the CX10 well (see Methods). The number of recovered MAGs and their estimated quality varies substantially between binning methods and using the same method with different parameter settings (**Table 1**). We identified 23 sets of MAGs representing the same microbial population within the CX10 community and recovered by at least five of the eight binning methods. Identical MAGs were only reconstructed by the different binning methods in two instances, whereas the estimated completeness or contamination of the MAGs differed by ≥5% for at least one binning method in 20 cases (data not shown).

Table 1. MAGs Recovered from a Deep Aquifer Metagenome by Varying Binning Methods

Comp. (%)[#]	Cont. (%)[#]	BinSanity	GroopM v2	MaxBin (ms 107)	MaxBin (ms 40)	MetaBAT v2	MetaBAT (specific)	MetaBAT (superspecific)	MetaBAT (verysensitive)
≥90	≤5	4	13	4	10	12	6	11	5
≥90	≤10	2	0	2	3	2	3	0	1
≥70	≤5	6	6	4	3	10	3	4	3
≥70	≤10	0	1	0	0	3	4	3	2
≥50	≤5	14	5	3	5	3	2	7	4
≥50	≤10	1	1	2	2	1	4	2	3
	Total	**27**	**26**	**15**	**23**	**31**	**22**	**27**	**18**

[#] Completeness (comp.) and contamination (cont.) of recovered MAGs was estimated using CheckM.

MAGs with an estimated completeness ≥50% and contamination ≤10% are considered to be of medium quality (Bowers et al., 2017). We further examined the relationship between medium-quality MAGs recovered by each binning method by placing these MAGs into a genome tree inferred from the concatenation of 23 ribosomal proteins (**Fig. 1a**). The unweighted UniFrac measure was then used to examine the phylogenetic β diversity between the different binning methods (**Fig. 1b**; Lozupone and Knight, 2005). Notably, the MAGs produced by MetaBAT and MaxBin under varying parameter settings were sufficiently different that these different applications of MetaBAT and MaxBin do not cluster together based of the pairwise β diversity measures (**Fig. 1b**). The deep aquifer microbial community is relatively simple as indicated by 75.4% of the metagenomic reads mapping to assembled contigs and recovered medium-quality MAGs accounting for 16.1% (MaxBin ms107) to 44.0% (MetaBAT v2) of assembled base pairs in contigs ≥1000 bp. Despite this community being amendable to binning, the different methods varied substantially in their results suggesting ensemble binning strategies may be able to outperform any individual method.

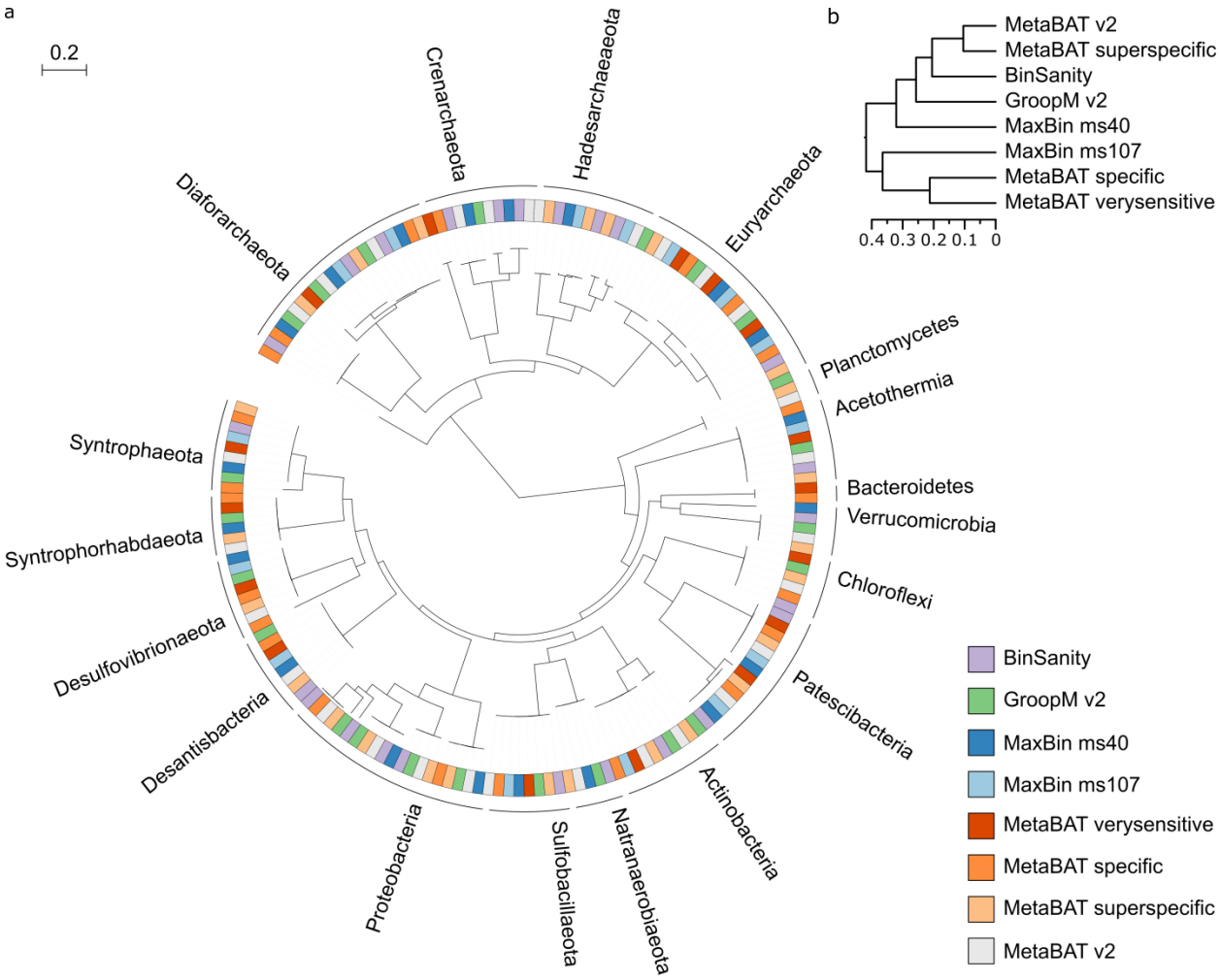


Figure 1. Comparison of MAGs recovered by different binning methods. (**a**) Distribution of MAGS produced by 8 binning methods on a tree inferred from the concatenation of 23 ribosomal proteins. A few MAGs were recovered by all 8 binning methods though the majority of MAGs were recovered by only a subset of the methods. (**b**) Phylogenetic beta diversity of the binning methods as determined using the unweighted UniFrac statistic and visualized as a UPGMA hierarchical cluster tree.

Ensemble Binning Workflow and Strategies

Ensemble binning produces a refined set of MAGs by integrating the results of multiple binning methods (**Fig. 2**). The greedy, consensus, and unanimous ensemble binning strategies proposed here can be applied to any number of binning methods and UniteM provides functionality for generating MAGs with several popular binning methods. All three strategies exploit the ability to estimate the completeness and contamination of MAGs in order to produce a refined set of MAGs. UniteM estimates the completeness and contamination of a MAG using the domain-

specific marker sets in CheckM (Parks et al., 2015) and defines the overall quality of a MAG as completeness - 2×contamination in order to preferentially favor MAGs with less contamination.

The three ensemble binning strategies operate in an iterative fashion to create a non-redundant set of MAGs from the redundant set of MAGs produced by different binning methods. The greedy strategy selects the highest quality MAG across all binning methods. Contigs comprising this MAG are then removed from all remaining MAGs in order to ensure the set of selected MAGs is non-redundant. Quality estimates for all remaining MAGs are then recalculated and this procedure repeated until the highest quality MAG falls below a user defined threshold. The consensus strategy aims to improve upon the greedy selection process by identifying sets of MAGs which represent the same microbial population recovered by different binning methods. All contigs within this 'matched set' of MAGs are then examined in order to produce a single refined MAG where individual contigs are retained or removed based on how often they appear in this matched set and how often they appears in other MAGs.

The greedy and consensus strategies aim to produce a large set of high quality MAGs. In contrast, the goal of the unanimous strategy is to reconstruct MAGs with substantially reduced contamination and provide an assessment of the stability of a MAG across different binning methods. This includes evaluating which binning methods place specific contigs within a MAG which can be used to support assertions of organism-specific metabolic pathways. This strategy proceeds in a similar manner as the consensus strategy, except a contig is only retained in a reconstructed MAG if it is present in all MAGs comprising the matched set.

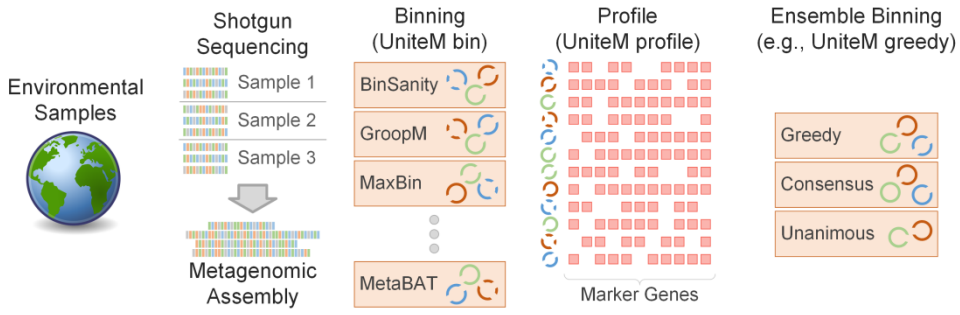


Figure 2. Recovery of MAGs using an ensemble binning strategy. Environmental samples are sequenced and assembled. MAGs are recovered using multiple different binning methods. UniteM provides "quality-of-life" functionality for producing MAGs from multiple popular binning methods with a single command. Marker genes used to estimate completeness and

contamination are then identified across all MAGs. UniteM identified marker genes as an independent step as this step is the most computationally expensive and is required by all downstream ensemble binning strategies. Ensemble binning strategies create a non-redundant set of MAGs from the redundant set of MAGs produced by different binning methods. The 'greedy' and 'consensus' strategies aim to produce a set of MAGs that is larger and of superior quality than provided by any individual binning method, while the 'unanimous' strategy aims to produce MAGs with reduced contamination.

Evaluation of Ensemble Binning Strategies using a Simulated Microbial Community

The high complexity simulated community generated for the Critical Assessment of Metagenome Interpretation (CAMI) competition was used to evaluate the proposed ensemble binning strategies (Sczyrba et al., 2017). This dataset simulates a 75 Gb time series dataset and consists of five samples taken from a synthetic community constructed from 596 genomes. BinSanity, GroopM v2, MaxBin, MetaBAT, and MetaBAT v2 were applied with varying parameter settings to the CAMI gold standard co-assembly and DAS Tool along with the greedy, consensus, and unanimous ensemble binning methods applied to the MAGs produced by these methods. We compared these approaches by considering the number of MAGs of at least medium quality (i.e., ≥50% completeness; ≤10% contamination) reconstructed by each approach (**Fig. 3**). The greedy, consensus, and DAS Tool ensemble binning strategies produced additional MAGs of higher quality than any individual binning method, though GroopM v2 and MetaBAT v2 perform extremely well on this dataset. DAS Tool and the greedy strategy had similar overall performance with the greedy strategy providing a few additional MAGs (482 vs. 475 medium-quality MAGs) with slightly increased completeness (435 vs. 423 MAGs ≥90%) and contamination (20 vs. 18 MAGs ≥5%). The refined MAGs produced by the consensus strategy resulted in 450 MAGs with a completeness ≥90%, the most produced by any approach, but also slightly increased contamination relative to the other ensemble methods (**Fig. 3b**). As expected, the unanimous strategy resulted in fewer recovered MAGs than other ensemble methods and even some individual binning methods. However, all 422 medium-quality MAGs provided by the unanimous strategy had zero contamination with the exception of two MAGs at 0.06% and 0.7% contamination.

Among the 596 genomes within the CAMI dataset are 240 genomes from species with multiple strains within the simulated community (Sczyrba et al., 2017). Binning methods often fail to produce MAGs of reasonable quality for populations with related strains within a sample

(Imelfort et al., 2014; Luo et al., 2015; Awad et al., 2017). The ensemble binning strategies are no exception with MAGs being recovered for 95.5 to 98.6% of the unique species, but only 35.0 to 56.3% of the common species (**Table 2**). In addition, the MAGs recovered from common strains were less complete and more contaminated, on average, than those from unique strains.

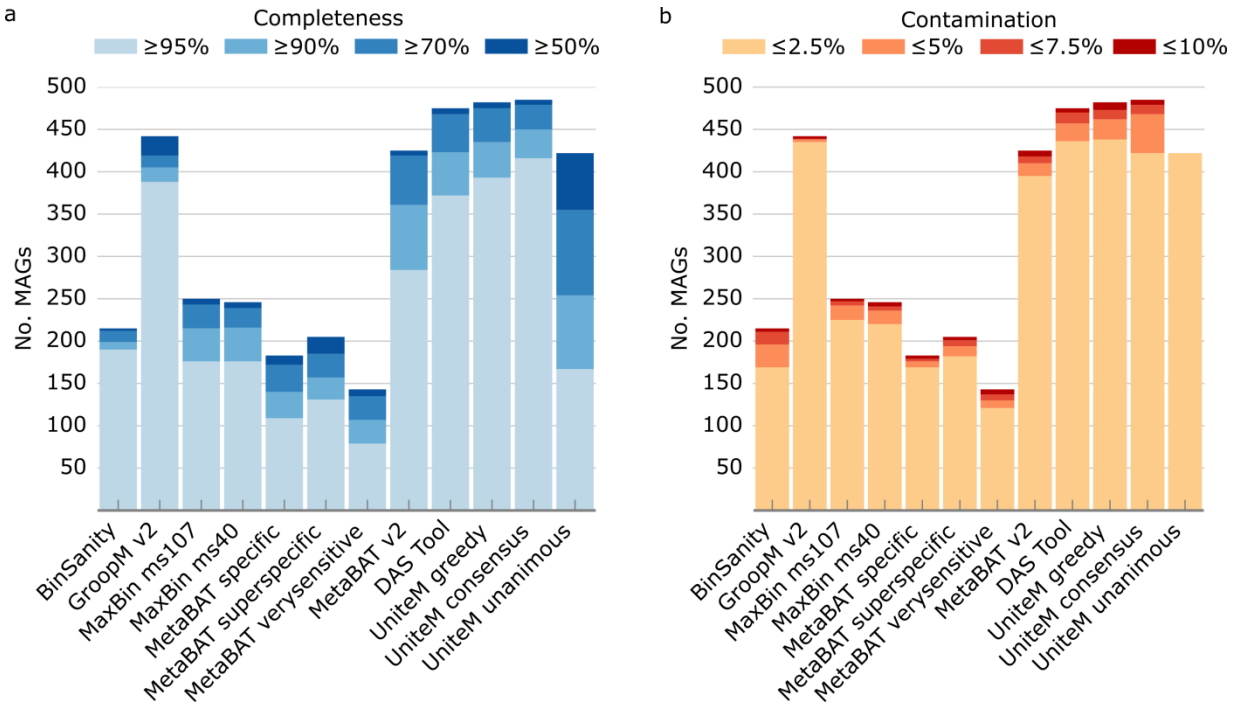


Figure 3. Performance of binning methods and ensemble binning strategies on a simulated microbial community consisting of 596 genomes. (**a**) Number of medium-quality MAGs with varying degrees of completeness. (**b**) Number of medium-quality MAGs with varying degrees of contamination.

Table 2. Recovery of Unique and Common Species in the CAMI Dataset

	No. Unique (of 356)	**Avg. Comp. (%)**	**Avg. Cont. (%)**	**No. Common (of 240)**	**Avg. Comp. (%)**	**Avg. Cont (%)**
GroopM v2	339	97.0	0.08	103	92.1	0.23
MetaBAT v2	340	95.0	0.41	85	93.8	1.14
DAS Tool	347	96.7	0.58	128	94.4	0.84
UniteM Greedy	347	97.6	0.59	135	93.7	1.13
UniteM Consensus	350	98.1	0.76	135	94.1	1.06
UniteM Unanimous	339	86.5	0.00	83	88.8	0.01

Ensemble Binning of a Deep Aquifer Metagenome

The greedy, consensus, and unanimous ensemble binning strategies were applied to the CX10 deep aquifer metagenome shown to have complementary MAGs recovered by different binning methods (**Fig. 1**; **Table 1**). Here we considered MAGs produced by these three strategies when applied with eight binning methods and a subset of three methods in order to evaluate the impact of using varying binning methods. For the subset of methods, MetaBAT specific, MetaBAT v2, and GroopM v2 were used as these methods produced large numbers of complementary MAGs (**Fig. 1b**; **Table 1**). The MAGs produced by the different binning and ensemble approaches were compared by estimating their quality with the lineage-specific marker sets of CheckM (Parks et al., 2015).

All ensemble binning strategies, including DAS Tool, produced additional MAGs of higher quality than any individual binning method when applied with only three or all eight binning methods (**Fig. 4**). The ensemble strategies selected MAGs across all eight binning methods (**Table 3**) and recovered one to three (2.7% to 8.3%) additional MAGs when considering all eight, instead of just three, binning methods. The best performing individual method was MetaBAT v2 which recovered 31 medium-quality MAGs compared to the 39 medium-quality MAGs produced by the greedy and consensus strategies. While the greedy strategy outperformed DAS Tool (39 vs. 36 medium-quality MAGs), the three additional MAGs recovered by the greedy strategy were all estimated to be ≤90% complete. These results may also exhibit some bias in favour of the greedy strategy as it uses the domain-level marker sets of CheckM to estimate MAG quality and the presented results are based on the lineage-specific CheckM quality estimates. The greedy and consensus strategies recovered the same number of MAGs, though similar to the results on the simulated microbial community the consensus strategy produced slightly more complete MAGs (total completeness of 2973 vs. 2923 across all 39 MAGs) at the expense of a slight increase in contamination (87 vs. 80). As expected, the unanimous strategy produced less complete MAGs when considering additional binning methods, but these MAGs exhibited less contamination.

We assessed the similarity of recovered MAGs by finding the most similar MAG, in terms of shared base pairs, produced by the DAS Tool, greedy, and unanimous strategies to each of the 39 medium-quality MAGs recovered by the consensus strategy (**Table 4**; **Supp. Table 1**).

Unsurprisingly, MAGs recovered by the unanimous strategy deviate substantially from the 'consensus' MAGs due to the filtering of contigs that were not unanimous across matched sets. While 23 of the MAGs recovered by the greedy strategy were nearly identical to 'consensus' MAGs, there were six MAGs which had <90% of their base pairs in common with any 'consensus' MAG. MAGs recovered by DAS Tool deviate even more substantially from the 'consensus' MAGs with only 14 being nearly identical to 'consensus' MAGs and 13 having <90% of their base pairs in common with a 'consensus' MAG. We directly compared the MAGs recovered by the greedy and DAS Tool strategies as these two approaches differ primarily in their methodologies for determining and defining genome quality. Despite their similarity, 12 of the 39 (30.8%) medium-quality 'greedy' MAGs had <95% of base pairs in common with a 'DAS tool' MAG (**Supp. Table 2**).

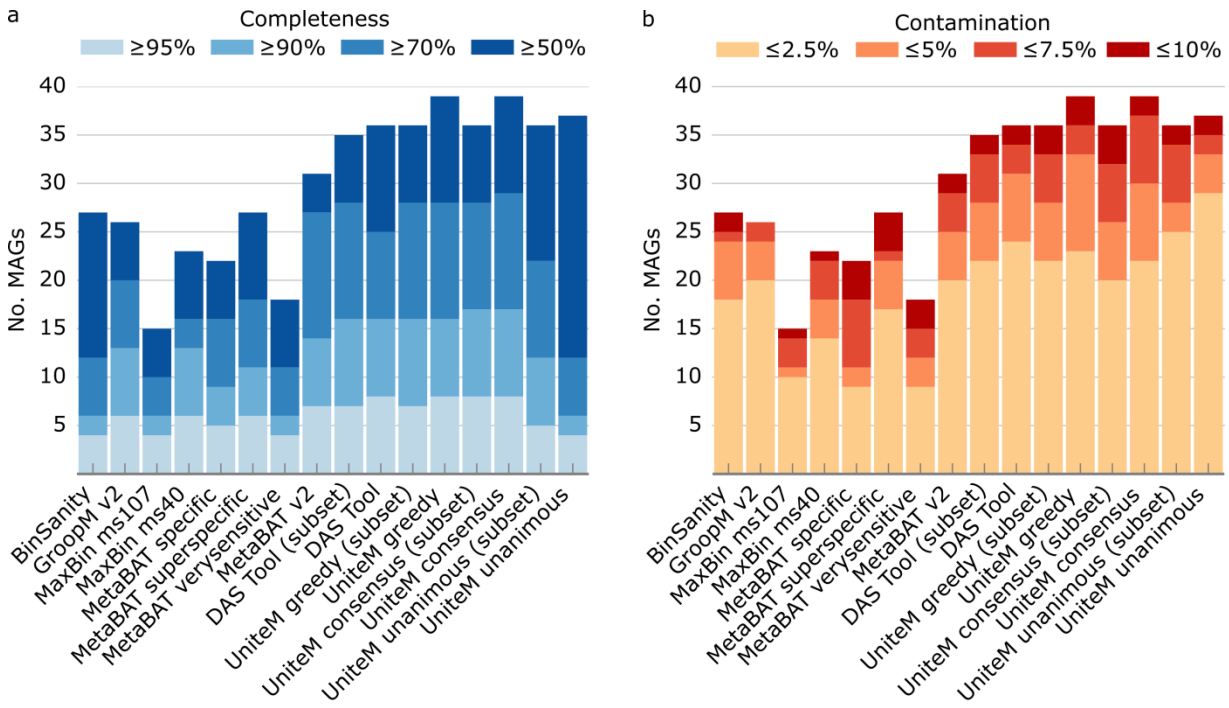


Figure 4. Performance of binning methods and ensemble binning strategies on the CX10 deep aquifer metagenome. (**a**) Number of medium-quality MAGs with varying degrees of completeness. (**b**) Number of medium-quality MAGs with varying degrees of contamination. Results are shown with the ensemble binning strategies applied to the MAGs produced by all eight binning methods and a subset of only three binning methods.

Table 3. Selection of Medium-quality MAGs across Binning Methods

Binning Method	Greedy	Consensus	Unanimous

BinSanity	6	6	5
GroopM v2	6	7	5
MaxBin ms107	1	1	1
MaxBin ms40	3	2	3
MetaBAT v2	7	7	7
MetaBAT specific	5	5	6
MetaBAT superspecific	2	3	2
MetaBAT verysensitive	9	8	8
Total	**39**	**39**	**37**

Table 4. Similarity of MAGs to 39 Medium-quality 'Consensus' MAGs

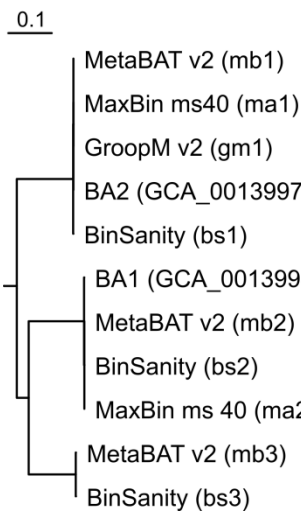
Shared Bases	DAS Tool	Greedy	Unanimous
≥99%	14	23	7
≥95%	7	5	3
≥90%	5	5	1
≥70%	9	6	13
<70%	4	0	15

Using the Unanimous Strategy to assess MAG Stability across Binning Methods

The results of the unanimous strategy can be examined to gain confidence in the specific contigs comprising a MAG. Here we inspected two Bathyarchaeota MAGs, BA1 and BA2, previously recovered from the CX10 deep aquifer metagenome (Evans et al., 2015). These MAGs are of interest as they are the first reported genomes outside the Euryarchaeota to encode the methyl-coenzyme M reductase complex (MCR), a complex diagnostic of archaeal methane/alkane metabolism (Evans et al., 2015; Laso-Pérez et al., 2016). MAGs for the BA2 population of at least medium quality were recovered by four of the binning methods, with the other four methods placing the BA2 contigs in chimeric MAGs (**Fig. 5**; **Supp. Table 2**). The unanimous strategy reconstructs a single MAG (88.5% complete; 4.7% contamination) consisting of only those contigs present in all four of the BA2 MAGs. This MAG shares 1256 of its 1467 (85.6%) genes with our previously reported BA2 MAG (Evans et al., 2015), including the MCR complex and other expected methane/alkane metabolism associated genes.

The BA1 population was more recalcitrant to recovery with only MetaBAT v2 (mb2), BinSanity (bs2), and MaxBin ms40 (ma2) producing at least medium-quality MAGs for this population (**Fig. 5**; **Supp. Table 3**). The MAG reconstructed by the unanimous strategy across these three MAGs was 91.5% complete and 0.9% contamination, but lacked the BA1 MCR complex as it

was absent from the mb2 and bs2 MAGs. However, this MAG shares 1739 of its 2340 (74.3%) genes with the previously reported BA1 MAG (Evans et al., 2015) and contains many of the other expected methane/alkane metabolism associated genes. Surprisingly, BinSanity placed the contig containing the BA1 MCR complex in another MAG (bs3) that resides in a lineage sister to the BA1 MAGs (**Fig. 5**). Examination of the bs3 MAG suggests this MCR-containing contig has been erroneously assigned to this MAG (**Fig. 6b**). Manually placing the MCR-containing contig into the mb2 and bs2 MAGs suggests this contig should be associated with these BA1 MAGs (**Fig. 6c, d**).



	Comp. (%)	Cont. (%)	No. Genes	OF (%)	MCR
MetaBAT v2 (mb1)	95.0	8.4	2193	84.8	BA2
MaxBin ms40 (ma1)	92.2	5.6	2249	85.9	BA2
GroopM v2 (gm1)	90.3	4.7	1715	81.6	BA2
BA2 (GCA_001399795.1)	93.8	3.7	1761	N/A	BA2
BinSanity (bs1)	90.7	9.0	1716	74.9	BA2
BA1 (GCA_001399805.1)	91.6	2.8	2403	N/A	BA1
MetaBAT v2 (mb2)	93.4	7.5	2728	75.3	none
BinSanity (bs2)	92.5	5.6	3052	75.7	none
MaxBin ms 40 (ma2)	92.5	4.7	2833	76.1	BA1
MetaBAT v2 (mb3)	78.7	5.6	2023	4.1	none
BinSanity (bs3)	78.5	4.7	1281	2.7	BA1

Figure 5. Bathyarchaeotal MAGs recovered by different binning methods. Phylogeny inferred from the concatenation of 23 ribosomal proteins containing the nine bathyarchaotal MAGs recovered by the eight binning methods along with two previously obtained MAGs, BA1 and BA2. The completeness (comp.), contamination (cont.), number of genes (no. genes), orthologous fraction (OF), and presence or absence of the BA1 or BA2 MCR complex (MCR) are given in the table. The orthologous fraction indicates the number of shared genes with either the BA1 or BA2 MAG.

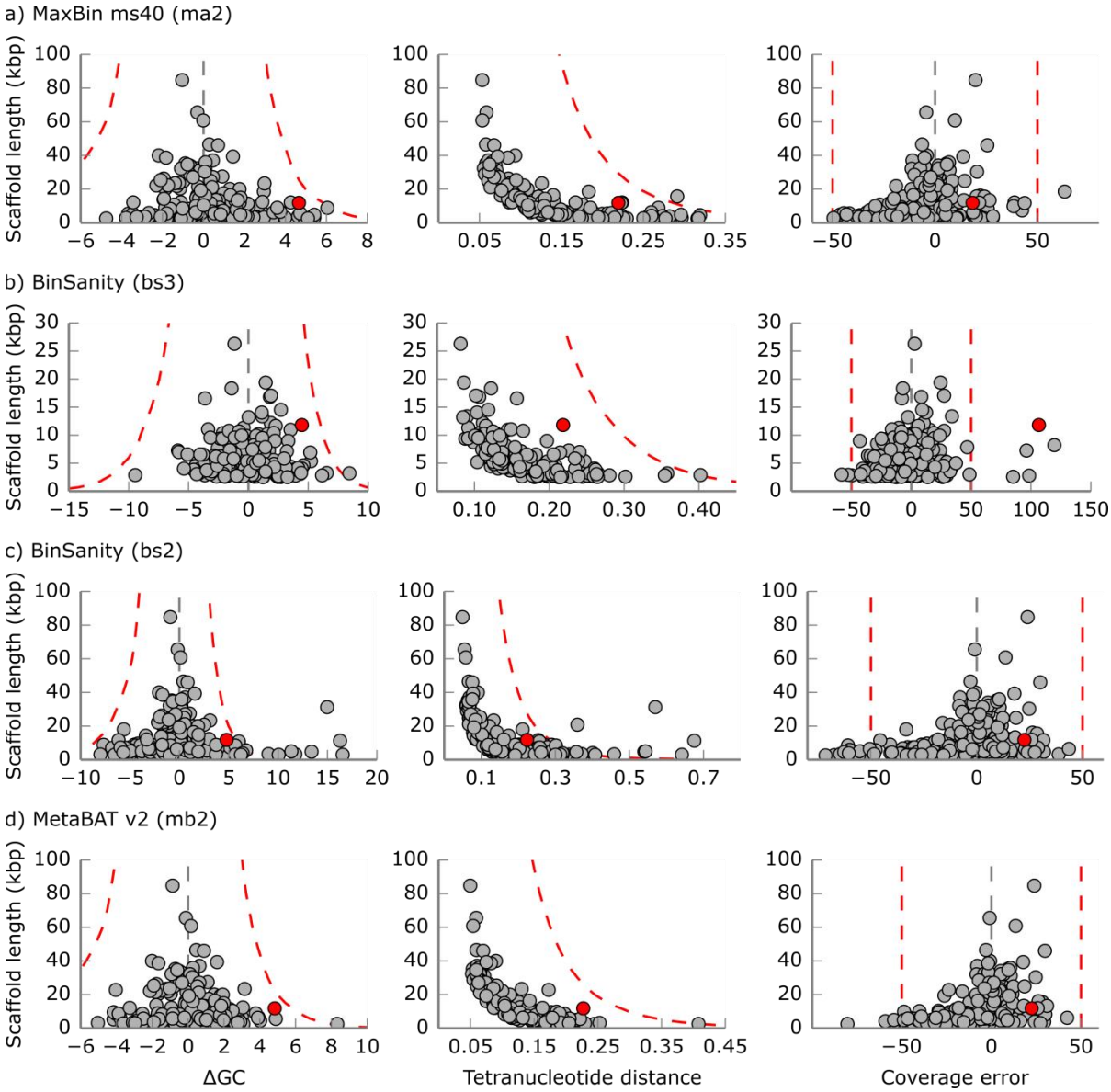


Figure 6. GC, tetranucleotide, and coverage distribution of bathyarchaeotal MAGs. Each point in the scatterplots represents a scaffold plotted as a function of its length (y-axis) and deviation from the MAGs mean GC-content, tetranucleotide signature, or coverage (x-axis). Dashed red lines represent the 98th percentile for the expected deviation in GC-content and tetranucleotide distances, or ±50% deviation in coverage. The scaffold in red contains the BA1 MCR complex. This contig was clustered with the ma2 and bs3 MAGs by their respective binning methods and manually assigned to the bs2 and mb2 MAGs.

Supp. Table 2. BA2 MAGs Produced by Different Binning Methods.

Binning Method	Completeness (%)	Contamination (%)	Classification
BinSanity	90.7	9.0	Medium quality
GroopM v2	90.3	4.7	Near complete
MaxBin ms107	100.0	61.8	Chimeric

MaxBin ms40	92.2	5.6	Medium quality
MetaBAT v2	95.0	8.4	Medium quality
MetaBAT specific	94.1	29.9	Chimeric
MetaBAT superspecific	95.0	12.2	Chimeric
MetaBAT verysensitive	95.3	36.0	Chimeric

Supp. Table 3. BA1 MAGs Produced by Different Binning Methods.

Binning Method	Completeness (%)	Contamination (%)	Classification
BinSanity	92.5	5.6	Medium quality
GroopM v2	86.5	30.4	Chimeric
MaxBin ms107	100.0	61.8	Chimeric
MaxBin ms40	92.5	4.7	Near complete
MetaBAT v2	93.4	7.5	Medium quality
MetaBAT specific	75.6	36.9	Chimeric
MetaBAT superspecific	94.4	47.1	Chimeric
MetaBAT verysensitive	100.0	145.0	Chimeric

Impact of Parameters on Ensemble Binning Strategies

The proposed ensemble binning strategies progress in an iterative fashion based on the estimated quality of MAGs across all binning methods. UniteM defines MAG quality as completeness - c×contamination, where c = 2 by default in order to favor MAGs with relatively low contamination. Here we determine the sensitivity of the ensemble binning strategies to how genome quality is defined by examining the MAG recovered from the CX10 deep aquifer metagenome when c is set to 1, 2, 3, 4, and 5 (**Fig. 7**). All three strategies show a reduction in the number of medium-quality MAGs recovered as c increases beyond two. This reduction in medium-quality MAGs is the result of preferentially selecting MAGs with increasingly lower levels of contamination at the expense of less completeness which ultimately results in the selection of MAGs that are <50% complete. However, this also has the expected benefit of generally reducing the number of MAGs with higher levels of contamination as c increases. We have set the default value of c to two in order to select MAGs with sufficient completeness to be of use in many downstream applications while penalizing MAGs with high levels of contamination, but researcher are free to use a larger value of c if they wish to further penalize contamination.

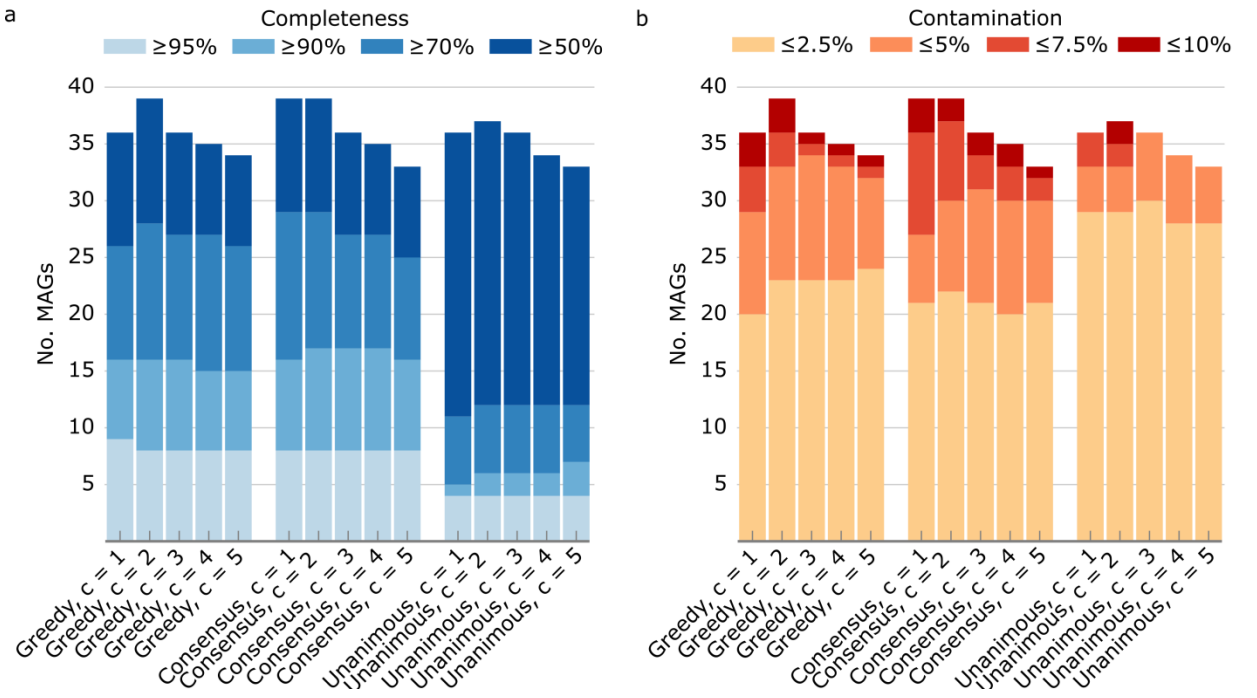


Figure 7. Performance of ensemble binning strategies with varying definitions of genome quality. Genome quality was define as completeness - *c*×contamination and results are shown for *c*={1,2,3,4,5}. (**a**) Number of medium-quality MAGs with varying degrees of completeness. (**b**) Number of medium-quality MAGs with varying degrees of contamination.

Ensemble Binning of 1,550 Public Metagenomes

We recently reported on the recovery of 7,903 MAGs from 1,550 public metagenomic samples (Parks et al., 2017). These MAGs were recovered by determining the optimal MetaBAT preset parameter setting for each individual metagenome. Here we explore the number of additional MAGs that can be obtained by applying the greedy and consensus ensemble binning strategies to MAGs produced by BinSanity, GroopM v2, MaxBin, MetaBAT, and MetaBAT v2. In order to compare results, we applied the same assembly filtering criteria to MAGs as in our previous study (e.g., N50 ≥10kb; Parks et al., 2017) and considered the number of MAGs defined as near complete (≥90% complete; ≤5% contamination), medium quality (≥70% complete; completeness - 5×contamination ≥50), or partial (≥50% complete; completeness - 5×contamination ≥50). The binning methods produce complementary results as indicated by all methods producing the largest number of MAGs of varying quality for at least some metagenomes and the best single overall method, MetaBAT v2, recovering the largest number of near complete MAGs for <50% of the metagenomes (**Supp. Table 4**).

Supp. Table 4. Number of metagenomes for which binning methods produces the largest number of MAGs of varying quality. Only metagenomes producing at least two MAGs of a given quality with the best performing binning method were considered. Methods tied for producing the largest number of MAGs for a metagenome were both credited.

Binning Method	Near Complete	Medium	Partial
BinSanity	42 (5.4%)	70 (6.1%)	74 (5.7%)
GroopM v2	272 (35.1%)	318 (27.9%)	300 (23.3%)
MaxBin ms107	226 (29.2%)	219 (19.2%)	185 (14.4%)
MaxBin ms40	286 (36.9%)	289 (25.4%)	273 (21.2%)
MetaBAT v2	381 (49.2%)	848 (74.4%)	988 (76.7%)
MetaBAT specific	327 (42.4%)	519 (45.5%)	583 (45.3%)
MetaBAT superspecific	369 (47.6%)	637 (55.9%)	692 (53.7%)
MetaBAT verysensitive	287 (37.0%)	438 (38.4%)	479 (37.2%)
Total Metagenomes	**775**	**1140**	**1288**

Table 5. MAGs Recovered from 1,550 Metagenomes

Binning Method	Near Complete	Medium	Partial	Total
Original Study (Parks et al. 2017)				
MetaBAT v2				
Greedy				
Consensus				

Discussion

Here we presented three ensemble binning strategies for combining the MAGs produced by different binning methods. The greedy strategy is perhaps the most intuitive and easiest to apply as the number and quality of MAGs will continue to improve with additional complementary binning methods (**Fig. 4**; Sieber et al., 2017). Despite the greedy strategy and DAS Tool selecting MAGs in a similar manner, the results of these methods vary as UniteM uses a larger set of marker genes for assessing the quality of MAGs as this will generally resulting in more robust estimates (Parks et al., 2015; **Table 4**; **Supp. Table 2**). The consensus and unanimous strategies produce refined MAGs by looking for agreement across MAGs from different binning methods that representing the same microbial population. As such, the quality of MAGs produced by these strategies is dependent on the binning methods considered. UniteM allows different subsets of binning methods to be easily considered in order to permit investigation of

how these subsets impact results. While the consensus strategy can be more challenging to apply than the greedy method, it can results in improved MAGs (**Fig. 3**; **Supp. Table 1**).

The main aim of the unanimous strategy is to provide an assessment of the stability of MAGs across different binning methods as illustrated here on two previously recovered bathyarchaeotal MAGs, BA1 and BA2 (Evans et al., 2015). The BA2 population exemplifies a population that has a core set of contigs that are grouped together by multiple binning methods. This core set supports the finding of a non-euryarchaeotal organism with an MCR complex and demonstrates that this result is independently predicted by multiple binning methods. The BA2 population also illustrates the benefits of considering multiple binning methods as the BA2 contigs were placed in chimeric MAGs by four of the eight methods. Notably, 211 of the 1467 (14.4%) core genes were not present in the previously reported BA2 MAG (Evans et al., 2015). The identification of new genes is not surprising given that the 'unanimous' BA2 and previously obtained BA2 MAGs are incomplete (88.5% and 93.8%) and contaminated (4.7% and 3.7%), and a different assembly algorithm was used in this study to assemble the CX10 metagenome (metaSPAdes as opposed to CLC). We stress that care must be exercised when interpreting core gene sets recovered by the unanimous strategy. While core genes are clustered together by multiple binning methods, these methods are not independent from each other and are subject to systematic errors in the assignment of contigs. For example, the 'unanimous' BA2 MAG is estimated to be 3.7% contaminated and there is little reason to believe this is purely the result of legitimate duplication of the marker genes used to estimate contamination.

Our results on the BA1 population illustrate the limitations of current binning methods and the benefits of manual refinement. Only three of the eight binning methods recovered a MAG of at least medium-quality for the BA1 population and only one of these contained the contig containing the BA1 MCR complex. However, manual inspection suggests that this MCR-containing contig should have been placed in the other BA1 MAGs and was erroneously binned by BinSanity (**Fig. 6**). These relatively poor binning results may be due to the BA1 population being absent or at extremely low abundance (<1x estimated coverage) in the other 10 aquifer metagenomes used to provide the binning methods with a differential coverage signal for the CX10 contigs. Ensemble binning strategies are limited by the quality of the MAGs provided by different binning methods. In the case of BA1, both the greedy and consensus strategies produce

a MAG that lacks the MCR complex as these strategies selected/refined the likely erroneous BinSanity bs2 MAG (**Fig. 5**).

A legitimate concern of ensemble binning strategies is that they will bias subsequent estimates of genome quality. To minimize this bias, UniteM uses domain-specific marker sets to estimate genome quality in order to enable more accurate lineage-specific marker sets to be used for subsequent quality estimates. While this helps to minimize bias in quality estimates, it does not eliminate it entirely as the quality of some MAGs will necessarily be estimated with domain-specific marker sets and there can be substantial gene overlap between domain- and linage-specific marker sets. Bias in quality estimates derived from marker genes is a growing concern as some binning methods make use of marker genes to help guide the binning process (Wu et al., 2016).

MAGs are providing the first opportunity to genomically exploration microbial communities, but this exploration is hampered by the quality of recovered MAGs and the presence of populations' recalcitrant to binning. The ensemble binning strategies presented here allow additional MAGs of improved quality to be obtained from microbial communities and can be used to assess the robustness of key findings. Until binning methods mature we expect ensemble strategies will provide a powerful tool for more fully understanding microbial communities.

Material and Methods

Simulate Microbial Community

The CAMI high complexity simulated community is a 75 Gb time series dataset with five samples taken from a synthetic community consisting of 596 genomes (Sczyrba et al., 2017). The simulated reads and gold standard assembly were obtained from the CAMI website (http://data.cami-challenge.org). Reads were mapped to the gold standard assembly with bwa v0.7.12 using the BWA-MEM algorithm and default parameters (Li and Durbin, 2009). Binning results were compared to the gold standard assembly to determine the completeness and contamination of each recovered MAG. Specifically, each MAG was assigned to the gold standard genome for which it shared the largest number of base pairs. Completeness was then calculated as the percentage of base pairs of the gold standard genome present in the MAG and

contamination as the percentage of base pairs within the MAG not belonging to the gold standard genome.

Deep Aquifer Metagenomes

Formation water from 11 coalbed methane wells in the Surat Basin (Queensland, Australia) were sampled in May or November of 2013 as previously described (Evans et al., 2015). The 100 bp paired-end Illumina HiSeq 2000 reads for these samples is available from the Sequence Read Achieve at the National Center for Biotechnology Information. Sequencing generated an average of 4.1±0.6 Gb of paired-end data for each sample. Trimmomatic v0.33 (Bolger et al., 2014) was used to remove adapters and quality trim reads with a sliding window threshold of Q15 and a leading/trailing threshold of Q3. Trimmed reads were merged using BBMerge v5.5 with default parameter settings (https://sourceforge.net/projects/bbmap/). Reads for the Coxon Creek 10 (CX10) sample were assembled using the metaSPAdes assembler of SPAdes v3.9.0 (Nurk et al., 2017) with default parameter setting. A differential coverage profile for the CX10 assembly was then obtained by mapping the reads for each of the 11 samples against this assembly with the BWA-MEM algorithm of bwa and default parameters.

Dataset of 1,550 Public Metagenomes

The 1,550 public metagenomes were obtained from the Sequence Read Archive (Leinonen et al., 2011). Information regarding these metagenomes and the methods used to assembly and determine coverage information was described previously (Parks et al., 2017). The assembled contigs and coverage information is ~350 GB and is available upon request from the authors.

Application of Binning Methods

MAGs for the simulated microbial community, CX10 deep aquifer metagenome, and 1,550 public metagenomes were obtained using BinSanity v0.2.5.9 (Graham et al., 2017), GroopM v2.0 (Imelfort et al., 2014), MaxBin v2.2.2 (Wu et al., 2016), MetaBAT v0.32.4 (Kang et al., 2015), and MetaBAT v2.10.2. BinSanity, GroopM, and MetaBAT v2 were run with default parameter settings; MaxBin was run with both the 40 and 107 marker gene sets; and MetaBAT v1 was run using the ‘verysensitive’, ‘specific’, and ‘superspecific’ preset parameter settings. DAS Tools v1.0 (Sieber et al., 2017) was run with default settings and BLASTP v2.6.0 (Camacho et al., 2009) or USEARCH v8.1.1861 (Edgar, 2010) to identify single copy gene in the simulated and deep aquifer metagenome, respectively.

Phylogenetic β Diversity of Binning Methods

A genome tree was inferred from the concatenation of 23 ribosomal proteins using FastTree v2.1.7 (Price *et al.*, 2009) under the WAG+GAMMA models as previously described (Parks et al., 2017). Fourteen of the 189 medium-quality MAGs produced by the 8 binning methods had amino acids in <25% of aligned columns and were removed before tree inference. Phylogenetic β Diversity was determined with the unweighted UniFrac measure (Lozupone and Knight, 2005) as calculated with Express Beta Diversity v1.0.7 (Parks and Beiko, 2013) and the similarity of the different binning measures visualized using the unweighted paired group method with arithmetic mean (UPGMA) hierarchical clustering algorithm (Sokal and Michener, 1958).

Comparison of MAGs

Bathyarchaeotal MAGs recovered from the CX10 metagenome were compared to those previously reported by Evans et al. (2015). The fraction of genes in common between MAGs was determined using CompareM v0.0.23 (https://github.com/dparks1134/CompareM) with sequence similarity determined by BLASTP v2.6.0 (Camacho et al., 2009) and genes being matched only if they had a percent identity ≥97% and an alignment that covered ≥97% of the genes. The orthologous fraction indicates the percentage of genes shared between two genomes and is calculated as s_{AB} / *min*(g_A, g_B), where s_{AB} is the number of shared genes between genomes A and B, and g_x is the number of genes in genome X. GC, tetranucleotide, and coverage distributions were obtained with RefineM v0.0.19 as previously reported (Evans et al., 2015).

Assessing Genome Quality for Ensemble Strategies

Note: the results presented in this manuscript used CheckM to identify domain-specific marker genes used to assess the quality of genomes. As of UniteM v1.0.0, CheckM is no longer used and the marker set is established as described in the UniteM manual.

The ensemble strategies of UniteM estimate genome completeness and contamination using the domain-specific, collocated marker sets of CheckM v1.0.7. These consist of 104 bacterial markers organized into 58 sets and 149 archaeal markers organized into 107 sets. MAGs are evaluated with both marker sets and assigned the completeness and contamination estimates with the largest sum. MAGs are assigned an overall quality defined as:

$$\text{completeness} - c \times \text{contamination} \quad (1)$$

where c = 2 by default in order to favor MAGs with low contamination.

Greedy Strategy

The greedy strategy operates in an iterative fashion to create a non-redundant set of MAGs from the redundant set of MAGs produced by different binning methods. The quality of each MAG is estimated using Equation 1 and the MAG with the highest estimated quality selected. If two or more MAGs have the same estimated quality, the MAG with the largest N50 statistic is selected with remaining ties resolved by selecting the largest MAG in terms of base pairs. Remaining ties are broken randomly. Contigs comprising the selected MAG are then removed from all other MAGs and the selected MAG removed from further consideration. This procedure is then repeat, starting with a re-assessment of MAG quality, until the highest quality MAG is below a defined quality threshold (default = 10).

This strategy is similar to the method used in DAS Tool, but differs in the methods used to assess the quality of MAGs.

Consensus Strategy

The consensus strategy removes MAGs of poor quality (by default, defined as having a quality <50%, completeness <50%, or contamination >10%) in order to consider only those MAGs that are a reliable indicator of which contigs should be binned together. This filtering is essential as the consensus strategy aims to produce refined MAGs by looking for agreement between MAGs recovered by different binning methods that represent the same microbial population. In descending order of MAG quality (Equation 1), we classifying any MAG that has >50% of its base pairs in common as being from the same population and refer to the set of MAGs satisfying this criterion as a matched set. This procedure is repeated until all MAGs have been assigned to a matched set, even if the set consists of a single MAG. The percentage of base pairs in common between two MAGs is calculated as follows where s is the number of shared bases, b_1 is the number of bases in the first MAG, and b_2 is the number of bases in the second MAG:

$$s \times 100 / \max(b_1, b_2) \quad (2)$$

Similar to the greedy strategy, the consensus strategy operates in an iterative fashion to create a non-redundant set of MAGs from the redundant set of MAGs produced by different binning

methods. This procedure starts by estimating the quality of each MAG quality (Equation 1) and organizing MAGs into matched sets. The cumulative quality of all MAGs in a matched set is then determined in order to identify the highest quality matched set, with ties resolved by considering the total N50 and genome size of the MAGs within each matched set. Remaining ties are broken randomly. MAGs within the highest quality matched set are then examined in order to create a single refined MAG by adding or removing contigs from the highest quality individual MAG within the set. A contig is removed from the highest quality MAG only if the contig is present in non-matched MAGs in the majority of binning methods (default: >50%). A contig is added to the highest quality MAG only if the majority of binning methods in which the contig is placed in a MAG are within the matched set (default: ≥50%) and the matched sets spans a minimum number of binning methods (default: 3). Contigs comprising the refined MAG are removed from all other MAGs and this procedure repeat until the highest quality matched set is below a defined quality threshold (default = 10).

Unanimous Strategy

The 'unanimous' strategy is a conservative method that identifies subsets of contigs that are consistently clustered together by different binning methods. It proceeds in a similar manner as the consensus strategy (see above), except a contig is only retained in the highest quality MAG if it is present in all MAGs in the matched set and contigs are never added to the highest quality MAG.

Implementation and Code Availability

UniteM is implemented in Python and is dependent on CheckM, which makes use of Prodigal (Hyatt et al., 2012) and HMMER (Eddy 2011). Source code is released under the GNU General Public License v3.0 and is available on GitHub at https://github.com/dparks1134/UniteM. Reported results are for UniteM v0.0.18.

Acknowledgements

The authors would like to thank the CAMI contributors for making their test sets available to the public and the many contributors to the Sequence Read Archive. DHP is supported by the Australian Centre for Ecogenomics and GWT by a University of Queensland Vice Chancellor's Research Focused Fellowship.

References

Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. 2013. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnol* **31**, 533-538.

Awad S, Luiz I, Brown CT. 2017. Evaluating metagenome assembly on a simple defined community with many strain variants. *bioRxiv* **155358**, doi: https://doi.org/10.1101/155358.

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

Bowers et al. 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaeal. *Nature Biotechnol* **35**, 725-731.

Brown CT, et al. 2015. Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208-11.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, doi: 10.1186/1471-2105-10-421.

Eddy SR. 2011. Accelerated profile HMM searches. *PLoS Comp Biol* **7**, e1002195.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460-61.

Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO2. 2015. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, doi: 10.7717/peerj.1319.

Evans PN, Parks DH, Chadwick GL, Robbins SJ, Orphan VJ, Golding SD, Tyson GW. 2015. Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics. *Science* **350**, 434-8.

Fred AL, Jain AK. 2005. Combining multiple clusterings using evidence accumulation. *IEEE Trans Pattern Anal Mach Intell* **27**, 835-50.

Graham ED, Heidelberg JF, Tully BJ. 2017. BinSanity: unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation. *PeerJ* **5**, doi: 10.7717/peerj.3035.

Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, doi: 10.1186/1471-2105-11-119.

Imelfort M, Parks DH, Woodcroft BJ, Dennis P, Hugenholtz P, Tyson GW. 2014. GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ* **2**, http://dx.doi.org/10.7717/peerj.603.

Kang DD, Froula J, Egan E, Wang Z. 2015. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, https://dx.doi.org/10.7717/peerj.1165.

Laso-Pérez R, et al. 2016. Thermophilic archaea activate butane via alkyl-coenzyme M formation. *Nature* **539**, 396-401.

Leinonen R, Sugawara H, Shumway M. The Sequence Read Archive. *Nuc Acids Res* **39**, D19-D21 (2011).

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760.

Lozupone C, Knight R. 2005. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* **71**, 8228-35.

Luo C, Knight R, Siljander H, Knip M, Xavier RJ, Gevers D. 2015. ConStrains identifies microbial strains in metagenomic datasets. *Nat Biotechnol* **33**, 1045-52.

Nielsen HB, et al. 2014. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat Biotechnol* **32**, 822-8.

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. *Genome Res* **27**, 824-834.

Parks DH and Beiko RG. 2013. Measures of phylogenetic differentiation provide robust and complementary insights into microbial communities. *ISME J* **7**, 173-83.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**, 1043-1055.

Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiol* *<in press>*.

Price MN, Dehal PS, Arkin AP. 2009. Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* **26**, 1641–1650.

Sangwan N, Xia F, Gilbert JA. 2016. Recovering complete and draft population genomes from metagenome datasets. *Microbiome* **4,** doi: 10.1186/s40168-016-0154-5.

Sczyrba A, et al. 2017. Critical Assessment of Metagenome Interpretation − a benchmark of computational metagenomics software. *bioRxiv*, https://doi.org/10.1101/099127.

Sieber CMK, et al. 2-17. Recovery of genomes from metagenomes via a dereplication, aggregation, and scoring strategy. *bioRxiv*, https://doi.org/10.1101/107789.

Sokal R and Michener C. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin* **38**, 1409–38.

Song WZ, Thomas T. 2017. Binning_refiner: improving genome bins through the combination of different binning programs. *Bioinformatics*, 10.1093/bioinformatics/btx086.

Strehl A, Ghosh J. 2002. Cluster Ensembles --- A Knowledge Reuse Framework for Combining Multiple Partitions. *Journal of Machine Learning Research* **3**, 583-617.

Strous M, Kraft B, Bisdorf R, Tegetmeyer HE. 2012. The binning of metagenomic contigs for microbial physiology of mixed cultures. *Front Microbiol* **3**, doi: 10.3389/fmicb.2012.00410.

Tennessen K, *et al*. 2016. ProDeGe: a computational protocol for fully automated decontamination of genomes. *ISME J* **10**, 269-72.

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.

Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NC, Wilkins MJ, Hettich RL, Lipton MS, Williams KH, et al. 2012. Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* **337**, 1661-1665.

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

Yeoh YK, Sekiguchi Y, Parks DH, Hugenholtz P. 2016. Comparative Genomics of Candidate Phylum TM6 Suggests That Parasitism Is Widespread and Ancestral in This Lineage. *Mol Biol Evol* **33**, 915-27.