MetaWRAP - a flexible pipeline for genome-resolved metagenomic data analysis

ABSTRACT

**Background**

The study of microbiomes with whole-metagenomic shotgun sequencing enables analysis of uncultivated microbial populations that may have important roles in their environments. Extracting individual draft genomes (bins) facilitates metagenomic analysis at the single genome level. Software and pipelines for such analysis have become diverse and sophisticated, creating a significant burden for biologists to access and use them. Further, while bin extraction algorithms are rapidly improving, there is still a lack of tools for their analysis and visualization.

**Results**

To address these challenges, we present metaWRAP, a modular pipeline software for shotgun metagenomic data analysis. MetaWRAP deploys state-of-the-art software to handle shotgun metagenomic data processing starting from raw sequencing reads, and ending in metagenomic bins and their analysis. MetaWRAP is flexible enough to give investigators control over their analysis, while still being easy-to-install and easy-to-use. It includes powerful hybrid algorithms that leverage the strengths of a variety of algorithms to extract and refine high-quality bins from metagenomic data through bin consolidation and reassembly. MetaWRAP’s hybrid bin extraction outperforms not only individual binning approaches, but also other bin consolidation programs in both synthetic and real datasets. Finally, metaWRAP comes with numerous modules for the analysis of metagenomic bins, including taxonomy assignment, abundance estimation, functional annotation, and visualization.

**Conclusions**

MetaWRAP is an easy-to-use modular pipeline that automates the core tasks in metagenomic analysis, while also contributing significant improvements to the assembly and interpretation of high-quality metagenomic bins. The bin refinement and reassembly modules of metaWRAP consistently outperform other binning approaches. Each module of metaWRAP is also a standalone component, making it a flexible and versatile tool for tackling metagenomic shotgun sequencing data. MetaWRAP is open-source software available at <https://github.com/ursky/metaWRAP>.

BACKGROUND

Compared to 16S rRNA amplicon sequencing, the study of microbial communities through whole metagenomic (WMG) shotgun sequencing opens new avenues for the investigation of the taxonomic composition of microbiomes, and their metabolic potential[1-3]. This greatly improves the ability to interpret and predict functional interactions, antibiotic resistance, and population dynamics of microbiomes, which has applications in human microbiome health, waste treatment, agriculture, and many others[4-6].

WMG shotgun sequencing reads come from hundreds or thousands of different community members, creating a unique and difficult analysis and interpretation challenge[3, 7]. Software for WMG data analysis has grown in number and complexity, improving our ability to extract individual genomes from this data[8-12]. However, these tools are burdensome for biologists to work with. As the field of WMG expands, comprehensive and accessible software for unified analysis of metagenomic data is needed[7, 11].

Establishing a WMG analysis pipeline requires investigators to find the best currently available tools, install and configure them on a cluster, and address conflicting libraries and environmental variables. Then they must run each software and script one after another, while converting the outputs of each tool into the correct format to input into the next step[13, 14]. These challenges present a major burden to anyone attempting metagenomic analysis, especially for investigators without computational experience. This limits the progress of microbial genomics as a field[15]. Existing automated pipelines and cloud services lack modularity, do not give users control over the analysis, and lack functions for genome-resolved metagenomics[14, 16-18]. The easy-to-install and easy-to-use metaWRAP tackles many of these challenges by making state-of-the-art WMG analysis more accessible to microbiologists, while retaining modularity and giving them control of the pipeline.

One aspect of metagenomic analysis is extracting the genomes of community members through binning of metagenomic assemblies[19]. Genome-resolved metagenomics allows for inspection of the metabolic pathways of individual taxa, as well as compare microbiomes at the scale of individual taxa. While many sophisticated tools such as CONCOCT, MaxBin, and metaBAT have been developed to tackle this problem, this is still an actively improving field[9, 20-22]. The k-mer composition, codon usage, and other sequence properties are expected to be similar throughout a given prokaryotic genome and scaffolds from the same organism are also expected to have similar read coverages in any given sample. Most metagenomic binning tools extract bins by clustering together scaffolds that have similar sequence properties and similar read coverages across multiple samples[23, 24].

For a metagenomic bin to be considered the genome of a single taxa, it must cover a significant length of the true genome (have a high completion), and also not have sequences belonging to other organisms (have a low contamination). The completion and contamination of a bin can be estimated by finding and counting universal single-copy genes that they have[27, 28]. CheckM improves on this by checking for single-copy genes that a genome of the bin’s taxonomy is expected to have[29].

There are a variety of approaches to binning, and no single method extracts the best version of each bin in every case. Bin consolidation tools attempt to combine the strengths and minimize weaknesses of different approaches. DAS\_Tool predicts single-copy genes in all the provided bin, aggregates bins with overlapping genes, and extracts a more complete consensus bin from each aggregate[25]. This collapsing approach significantly improves the completion of the bins. Binning\_refiner, on the other hand, splits the contigs into bins such that all contig division boundaries of the original predicted bins are satisfied. This breaks the contigs into many more bins, reducing contamination[26]. These approaches consolidate sets of bins from different methods and result in a superior bin set, but they have limitations – DAS\_Tool increases completion at the expense of introducing contamination, while Binning\_refiner prioritizes purity, but loses completeness. MetaWRAP’s Bin\_refinement module addressed these limitations by incorporating both a splitting and a collapsing step.

A relatively unexplored way to improve draft genome quality is bin reassembly – extracting reads that belong to a given bin and assembling them separately from the rest of the metagenome. With proper benchmarking, this approach could significantly improve the quality and downstream functional annotation of at least some bins in a microbial community. This approach is implemented in the metaWRAP-Reassemble\_bins module.

Because the field is relatively new, there is a lack of software to inspect, analyze, and visualize metagenomic bins. While there are tools that can accurately predict the taxonomy of metagenomic scaffolds (such as Taxator-tk), there is no tool to classify entire metagenomic bins[30, 31]. Similarly, there are many ways to estimate the coverage of scaffolds based on read alignment depth, but no way to find the coverages of entire bins across many samples[32, 33]. Finally, there no tool to visualize draft genomes in context of whole metagenomic communities. These knowledge gaps inspired the construction of metaWRAP’s Quant\_bins, Classify\_bins, and Blobology modules.

RESULTS AND DISCUSSION

**MetaWRAP is a flexible, modular pipeline**

MetaWRAP is a modular, comprehensive platform for analysis, visualization, and interpretation of metagenomic data, with emphasis on extracting and analyzing high-quality draft genomes (bins). It is open-source software available at <https://github.com/ursky/metaWRAP>. With the help of Bioconda (<https://anaconda.org/ursky/metawrap-binning>), metaWRAP can be easily installed locally and on remote clusters[34]. The metaWRAP installation produces a bioinformatics environment with over 150 commonly used bioinformatics software and libraries (Figure S1). MetaWRAP itself is a collection of modules, each of which uses a variety of pre-existing software, newly developed software, and databases to accomplish each specific step of metagenomic analysis. Unlike existing metagenomic wrappers and cloud services, metaWRAP retains modularity and grants the user control of the analysis pipeline. The user may follow the intuitive workflow starting from raw metagenomic shotgun sequencing reads all the way to high-quality draft genomes and their functional annotation, or use only the functions they are interested in, as each module is also a standalone program (Figure 1).

The first metaWRAP modules pre-process metagenomic data by quality-controlling and taxonomically profiling the reads, and assembling them. MetaWRAP-Read\_qc module trims the raw sequence reads, removes human contamination, and produced quality reports for each of the sequenced samples. The reads from all given samples can be then assembled with the metaWRAP-Assembly module with MegaHit[35] or metaSPAdes[36], also producing an assembly report. Both the reads from each sample and the assembly can be taxonomically profiled with the Kraken[30] module, producing interactive kronagrams[37] of community taxonomy. The assembly is then binned with the metaWRAP-Binning module by three metagenomic binning software – MaxBin2, metaBAT2, and CONCOCT[19, 21, 22].

The other modules focus on refining, analyzing, and visualizing metagenomic bins from either the Binning module or other sources. MetaWRAP-Bin\_refinement module hybridizes to three bin sets with Binning\_refiner[26], and then finds the best version of each bin based on completion and contamination metrics estimated with CheckM[29] (Figure S2). The scaffolds in the final bin set are then de-replicated, and a report of their completion, contamination, and other metrics is produced. MetaWRAP-Reassemble\_bins can then be used to reassemble the reads belonging to each bin, improving their N50, completion, and contamination (Figure S3).

The resulting bins can then be visualized by using the metaWRAP-Blobology module[38], which plots the contigs of the joint assembly on a blob plot, annotating them with their taxonomy and bin membership. The metaWRAP-Quant\_bins module can be used to quickly estimate the abundance of each bin in each of the metagenomic samples. MetaWRAP-Classify\_bins can be used to conservatively, but accurately estimate their taxonomy. Finally, the bins can be functionally annotated with the metaWRAP-Annotate\_bins module.

**MetaWRAP-Bin\_refinement improved bin predictions in synthetic data**

To test the efficacy of the metaWRAP-Bin\_refinement module at consolidating and improving bin sets, we used synthetic metagenomic data sets of varying complexity from the Critical Assessment of Metagenomic Interpretation (CAMI) study[9]. The “gold standard” assemblies from the “high”, “medium”, and “low” diversity challenges were first binned with metaBAT2, Maxbin2, and CONCOCT using the metaWRAP-Binning module, and the resulting three bin sets were then consolidated with DAS\_Tool[25], Binning\_refiner, and metaWRAP-Bin\_refinement.

The completion and contamination of the bins in the original and refined bin sets were evaluated with CheckM (Figure S4) and Amber[39]. True recall and precision for each bin calculated with Amber (Figure S5) were converted to completion and contamination percentages to be comparable to the CheckM results (Figure 2). We found that metaBAT2 consistently outperformed MaxBin2 and CONCOCT, producing a total of 385 high quality bins between all the challenges (completion greater than 90% and contamination less than 5%), and 271 near-perfect bins (completion greater than 95% and contamination less than 1%). MaxBin2 came in second with 275 high quality bins and 164 near-perfect bins. CONCOCT performed rather poorly in all but the smallest CAMI challenge data sets, producing 58 high quality bins and 40 near-perfect bins.

In the consolidated bin sets, DAS\_Tool produced 426 high quality bins and 263 near-perfect bins across all CAMI challenges, while Binning\_refiner produced 289 and 210 bins, respectively. DAS\_Tool consistently produced high completion bins, however these bins had relatively high contamination, which is a result of the aggregation approach that DAS\_Tool takes. Binning\_refiner on the other hand produced very pure bins with its splitting approach, however it did so at the expense of significantly reduced completion. MetaWRAP-Bin\_refinement produced bins that had both high completion and low contamination. In total, it produced 457 high quality bins and 339 near-perfect bins (Figures 2, S4) due to both a splitting and aggregation step. These results confirm that metaWRAP not only consistently improves bin sets through its consolidation approach, but it also outperforms other consolidation algorithms in data sets of varying complexity.

The use of CheckM (Figure S4) and Amber (Figure 2) to evaluate the binning sets produced similar results, although overall CheckM slightly overestimated both completion and contamination of the produced bins. More importantly, however, the relative performance of the six binning approaches was the same when evaluating with CheckM or Amber. This validates the use of CheckM for benchmarking binning results in data sets where the true genomes remain unknown.

**Bin\_refinement improves bin predictions in real data**

MetaWRAP was also benchmarked against real metagenomic data, using water, gut, and soil microbiome WGS data sets. The water data set was from a brackish water survey of the Baltic Sea[40] and includes 36 samples for a total of 196Gbp of sequence. The gut data set came from the Metagenomic of the Human Intestinal Tract (MetaHIT) survey[41] and consisted of 50 samples and a total of 144Gbp. The soil data set came from grassland soil microbial communities from Angelo Coastal Reserve[25] and included 6 samples for a total of 481Gbp of sequencing data.

The samples from each microbiome type were pre-processed through the metaWRAP-Read\_qc module to trim reads and remove human contamination, and the Kraken module was used to obtain the taxonomic profile of the community (Figure S7). The water samples were dominated by *Alphaproteobacteria* and *Actinobacteria*, the gut samples were dominated by *Bacteroidetes* and *Clostridia*, and the soil samples comprised of a wide variety of *Proteobacteria* and *Terrabacteria.* (Figure 3)

The quality-controlled reads were then co-assembled with the metaWRAP-Assembly module and the assemblies binned with metaBAT2 Maxbin2, and CONCOCT using the metaWRAP-Binning module. The resulting three bin sets of each microbiome type were then consolidated with DAS\_Tool, Binning\_refiner, and metaWRAP-Bin\_refinement, and the completion and contamination of all the resulting bins were evaluated with CheckM. (Figure 3)

Between the original binning software, metaBAT2 consistently produced the best sets of bins when compared to MaxBin2 and CONCOCT, producing 202, 146, and 88 acceptable quality bins (comp ≥ 50%, cont ≤ 10%) in water, gut, and soil samples, respectively. MaxBin2 had 151, 98, and 40 bins, and CONCOCT 65, 121, and 39 bins.

Despite incorporating all the binning methods, DAS\_Tool struggled to improve the original bin sets, producing 198, 130, and 63 acceptable quality bins in water, gut, and soil samples, respectively. DAS\_Tool performed relatively well at higher bin completion ranges (≥ 80%), although at the expense of increased contamination. Binning\_refiner performed similarly, with 206, 138, and 83 bins in water, gut, and soil data sets, respectively. The bins from Binning\_refiner were less complete, but also had significantly lower contamination than bins in the original bin sets.

MetaWRAP’s Bin\_refinement produced 235, 175, and 134 acceptable quality bins in water, gut, and soil samples, respectively, significantly outperforming all other tested approaches. The module uses Binning\_refiner in its pipeline to hybridize the input bin sets, and then choses the best version of each bin from the original and hybridized sets. Because the Bin\_refinement module leverages the strength of Binning\_refiner but still has a collapsing step similar to DAS\_Tool, it is able to match DAS\_Tool’s high completion rankings, while retaining the low contamination rankings of Binning\_refiner. MetaWRAP consistently produced the highest quality bin sets in all the tested metagenomic data sets, which ranged greatly in diversity, taxonomic composition, and sequencing depths.

It is important to note that the use of metaWRAP’s Bin\_refinement module to improve binning predictions is not limited to the bin sets produced from the metaWRAP-Binning module (metaBAT2, MaxBin2, and CONCOCT). Bin sets from any 2 or 3 binning software may be used as input for the module. Furthermore, because the algorithm leverages the differences between the input bin predictions, it is also possible to use bin sets produced from different parameters of the same software as input, and the metaWRAP-Bin\_refinement will consolidate them into a superior bin set.

**Bin\_refinement adjusts to the desired bin quality**

To consolidate the original and hybridized bin sets, metaWRAP-Bin\_refinement chooses the best version of each bin based on their completion and contamination values. However, this selection is subjective, and depends on what the user believes to be the “best bin”. The minimum completion (-c) and maximum contamination (-x) options are key parameters that greatly alter the quality of the bins produced, as the module will dynamically adjust its algorithms to produce the maximum number of bins in this range.

To demonstrate the effects of changing the –c and –x parameters of metaWRAP’s Bin\_refinement module, we ran the original bin sets from water, gut, and soil data sets with varying minimum completion (but fixed maximum contamination), and varying maximum contamination (but fixed minimum completion) parameters. When compared to the original Bin\_refinement (-c 50 –x 10), the module produced a greater number of bins at any given threshold when it was given appropriate –c and –x parameters. (Figures S6, S7)

The improvements were especially noticeable at higher completion and lower contamination ranges. For example, MetaWRAP-Bin\_refinement with –c 90 –x 10 recovered 19, 18, and 1 (water, gut, and soil, respectively) extra bins with a minimum completion of 90%, when compared to the baseline –c 50 –x 10 run. Similarly, MetaWRAP-Bin\_refinement with –c 50 –x 1 parameters extracted 8, 21, and 4 (water, gut, and soil, respectively) more bins at a maximum contamination of 1%, when compared to the baseline run.

Unlike arbitrary and confusing thresholding parameters in many other software, the minimum completion and maximum contamination options offer the user an intuitive way to parameterize the metaWRAP’s Bin\_refinement module to their needs. This leads to significant increases in the number of quality bins they are able to extract from their data.

**Reassemble\_bins significantly improves bin quality**

MetaWRAP’s Reassemble\_bins module improves a given set of bins through individual reassembly with SPAdes[42]. The module only replaces the original bins if the reassembled ones are better than in terms of completion and contamination. Like the Bin\_refinement module, the Reassemble\_bins module takes in minimum completion (-c) and maximum contamination (-x) parameters to allow the user to define what they consider a “good” bin. The bins produced from the water, gut, and soil data with metaWRAP-Bin\_refinement module runs (–c 50 –x 10) were run through the metaWRAP-Reassemble\_bins module (-c 50 –x 10), and the resulting bins were re-evaluated with CheckM.

The Reassemble\_bins module was able to improve upon 78%, 98%, and 2% of the bins in the water, gut, and soil bin sets, respectively. The module significantly improved the water and bin set overall metrics, increasing their N50 and completion scores. Even more strikingly however, the reassembly process significantly reduced contamination in these bin sets. (Figure 5).

The success of the bin reassembly algorithm relies heavily on accurate and specific recruitment of the correct reads to each bin. In very diverse and heterogeneous communities such as those found in soil, the read recruitment may not be specific enough. This confuses the assembler during the re-assembly stage, and results in an improvement in only a small fraction of the bins. However, draft genomes from gut and water samples were still significantly improved with the Reassemble\_bins module despite their complexity (Figure 3).

**MetaWRAP produces high-quality draft genomes**

We investigated the performance of different binning approaches (both original binners and bin consolidation software) when extracting high quality draft genomes, with a contamination less than 5% and completion greater than 70%, 80%, 90%, and 95%. The default run of metaWRAP-Bin\_refinement consistently produced the highest number of high-quality draft genomes in water, gut, and soil data sets. These numbers further improved when re-running the module with appropriate minimum completion (-c) settings (i.e running Bin\_refinement –c 90 when benchmarking for bins with a minimum completion of 90%). This approach significantly outperformed every other tested binning and bin refinement method at every quality threshold.

The reassembly of the metaWRAP-derived bins with the Reassemble\_bins module made a further improvement on the number of high-quality draft genomes extracted from the gut and water data sets. Even the default run of Reassemble\_bins produced a significantly better bin set compared to non-reassembled bin sets produced by all tested software, including metaWRAP’s Bin\_refinement. However, just like in the Bin\_refinement runs, the results were even better when Reassemble\_bins was provided with an appropriate –c option.

When comparing to the original binning software (MaxBin2, metaBAT2, and CONCOCT) and bin consolidation tools (DAS\_Tool and Binning\_refiner), metaWRAP produced the largest number of high-quality draft genomes in all the tested WMG data sets. Additionally, it should also be considered that metaWRAP is capable of improving bin sets from any binning software. Therefore, even when better metagenomic binning software are developed, their outputs can still be further improved with metaWRAP refinement and reassembly algorithms.

**MetaWRAP enables analysis and visualization of metagenomic bins**

The rest of the modules address examining and processing a set of bins in preparation for downstream analysis. The user may visualize the bins in context of the entire community with the Blobology module, quantify their abundances across samples with the Quant\_bins module, estimate their taxonomy with the Classify\_bins module, and functionally annotate them with the Annotate\_bins module.

The metaWRAP-Quant\_bins module was used to estimate bin abundances across samples, and the results are shown in a clustered heatmap (Figure S9). Clustered heatmaps like these may be used to infer bin co-abundance, as well as identify composition relationships between samples. Because this approach considers the abundances of every extracted bin individually, it offers higher resolution information than looking at the community differences at higher taxonomic ranks.

Bins were also visualized with the metaWRAP-Blobology module. The module produces GC vs Abundance plots of contigs, annotated with their taxonomy[43] (Figures 3 and S11) or bin membership (Figure 7). These plots allow for inspection of the extracted bins in the context of the entire community that they belong to, as well as visualize the relative success of the binning process.

The final reassembled bins were taxonomy profiled with the metaWRAP-Classify\_bins module (Figure S10), functionally annotated with the Annotate\_bins module. Together, this information may be used in downstream analysis to investigate complex questions about functional interactions and metabolic potential of individual community members.

CONCLUSIONS

Genome-level analysis of whole genome metagenomic sequencing data is essential in understanding the composition and function of microbiomes. Until now, however, this rapidly growing field lacked a unifying platform to utilize the wealth of currently available software and make it easily accessible to researchers. MetaWRAP is a modular pipeline that can handle common tasks in WMG data analysis starting from raw read quality control, and ending in bin extraction and analysis. MetaWRAP is easy to install through Bioconda, simple to use, and if its modularity gives the investigator flexibility in their analysis approach.

MetaWRAP contributes significant improvements to the recovery of draft genomes from WMG data through bin refinement and reassembly. The bin refinement module uses a novel hybrid approach to consolidate bin predictions from different binning software, producing a single stronger set. This approach significantly outperforms individual binning software, as well as other consolidation algorithms. The algorithm can adjust to accommodate specific draft genome quality targets, making it suitable for many research applications. MetaWRAP’s bin reassembly module further improves the draft genomes in both completeness and purity. Finally, metaWRAP contains multiple modules for analysis and evaluation of metagenomic bins – bin taxonomy assignment, abundance estimation, functional annotation, and visualization.

METHODS

For detailed descriptions of each of metaWRAP’s modules, please refer to Supplementary Methods.

**CAMI binning benchmarking**

The “gold standard” assemblies from the “high”, “medium”, and “low” diversity CAMI challenges were binned with metaBAT2 v2.12.1[22], Maxbin2 v2.2.4[21], and CONCOCT v0.4.0[19] using the metaWRAP-Binning module with default parameters. The resulting three bin sets were then consolidated with DAS\_Tool v1.1.0[25] (default settings, blast used for search engine), Binning\_refiner v1.2[26] (default settings), and metaWRAP-Bin\_refinement v0.7 (see Supp. Methods for module details) to attempt to improve the bin sets. To simulate a realistic metagenomic pipeline, the completion and contamination of the bins in all six bin sets was first evaluated with CheckM v1.0.7[29] with default parameters, and bins with a completion less than 50% or a contamination greater than 10% were discarded. The true recall and precision of the bins within the six resulting bin sets was then determined with Amber v0.6.2[39], which compared the bins against the known original genomes. Bin recall and precision were converted to completion and contamination percentages.

**Real data binning benchmarking**

The raw sequences from water, gut, and soil microbiomes were run through the metaWRAP-Read\_qc module, which trims the reads with TrimGalore[44], removes human contamination with BMTagger[45] searching againsts hg38, and produces a quality report with FASTQC[46]. MetaWRAP’s Kraken module (-s 10000000) was then run on the quality-controlled reads to investigate the taxonomic profile of the community with Kraken[30] (using standard database) and KronaTools 2.7[37]. The reads were then co-assembled within each community type with MegaHit v1.1.2[35] by using the metaWRAP-Assembly module. Contigs shorter than 1000bp were discarded, with the exception of the soil assembly, for which the cutoff of 3000bp was chosen to reduce the binning time.

The co-assemblies of each data type were then binned with metaBAT2 v2.12.1, Maxbin2 v2.2.4, and CONCOCT v0.4.0 using the metaWRAP-Binning module at default settings. The resulting three bin sets of each microbiome type were then passed to DAS\_Tool v1.1.0 (--search\_engine blast option), Binning\_refiner v1.2 (default settings), and metaWRAP-Bin\_refinement v0.7 to attempt to improve the bin sets. For the main benchmark, metaWRAP was run with –c 50 –x 10 settings. To benchmark the bins produced by all the binning methods, the completion and contamination of the bins was estimated with CheckM v1.0.7. See Supplementary Methods for details on all modules.

**Bin\_refinement optimization demonstration**

The metaWRAP-Bin\_Refinement module was run with a variety of settings to demonstrate performance changes at different –c (minimum completion) and –x (maximum contamination) settings. First, the bin sets produced with metaBAT2 v2.12.1, Maxbin2 v2.2.4, and CONCOCT v0.4.0 were refined with the module with a constant maximum contamination setting –x 10, but varying minimum completion settings –c 50, 60, 70, 80, 90, and 95. Then the same bin sets were refined with a constant minimum contamination setting –c 50, but varying maximum contamination setting of –x 10, 8, 6, 4, 2, and 1.

**Reassembly benchmarking**

To benchmark overall reassembly performance, bin sets produced by the metaWRAP-Bin\_refinement module with -c 50 –x 10 settings were run through the metaWRAP-Reassemble\_bins (see Supp. Methods for module details) module with -c 50 –x 10 settings. The re-assembly module uses BWA 0.7.15[33] and Samtools 1.6[47] to pull reads belonging to each bin, and then reassembled them with SPAdes[42] (--carefull option). The resulting bins were evaluated with CheckM v1.0.7, and the completion and contamination values were sorted and plotted.

**Extracting high-quality draft genomes**

MetaWRAP’s Bin\_refinement and Reassemble\_bins modules were run with different settings to extract high quality draft genomes (contamination less than 5%, completion greater than 70%, 80%, 90%, or 95%) to showcase the overall binning potential of metaWRAP. To benchmark the Bin\_refinement module, it was run on bin sets produced with metaBAT2 v2.12.1, Maxbin2 v2.2.4, and CONCOCT v0.4.0 with four different settings: -c 70 –x 5, -c 80 –x 5, -c 90 –x 5, -c 95 –x 5. To benchmark the Reassemble\_bins module, it was run with the same settings on the output of of Bin\_refinement with -c 60 –x 10, -c 70 –x 10, -c 80 –x 10, and –c 90 –x 10 settings, respectively. Finally, all the resulting metaWRAP bin sets, the original bin sets, as well as the refinements from DAS\_Tool and Binning\_refiner were evaluated with CheckM v1.0.7 and the number of bins with different completion and contamination values were counted and plotted.

**Draft genomes analysis**

Bins produced with metaWRAP-Bin\_refinement (-c 70 –x 10 options) were then visualized with the Blobology module (--bins flag used to provide bins), which uses a modified Blobology[38] scripts, Bowtie2 2.3.2[48], and MegaBLAST 2.6.0[43] to make Taxon-Annotated-GC-Coverage plots. The abundance of these bins in each sample was then estimated and visualized with the Quant\_bins module, which uses Salmon 0.9.1[49] to quantify individual contigs[50] and then estimate bin abundances.

The reassembled bins from the metaWRAP-Reassemble\_bins module (-c 50 –x 10 options) were then run through the Classify\_bins module (default settings), which makes initial taxonomy predictions of individual scaffolds with Taxator-tk 1.3.3e[31] and then estimates the taxonomy of entire bins. These bins were then functionally annotated with the metaWRAP-Annotate\_bins module, which uses PROKKA 1.12[51] to annotate each bin in parallel.

OTHER SECTIONS

**Additional files**

1. Supplementary Methods
2. Figure S1: Detailed walkthrough of the data files, software, databases, and custom scripts that metaWRAP uses. The components of each metaWRAP module grouped and denoted with dotted lines.
3. Figure S2: Logical workflow of the Bin\_refinement modules of metaWRAP. The module takes in three bin sets produced from the same assembly by different software or different parameters of the same software. Binning\_refiner is used to create hybridized intermediates (4 possible combinations), and the completion and contamination of the original and hybridized bins is estimated with CheckM. The best version of each bin is then found in the resulting 7 bin sets.
4. Figure S3: Logical workflow of the Reassemble\_bins module, which extracts reads belonging to bins in a given bin set, and individually reassembles them. This process is done for perfectly mapping reads (strict) and reads mapping with less than 3 mismatches (permissive). For each bin, CheckM is used choose the best bin between the original and the two reassembled versions.
5. Figure S4: Completion and contamination (determined with CheckM) of bins recovered from the CAMI’s high, medium, and low complexity synthetic data sets using original binning software (metaBAT2, MaxBin2, CONCOCT) and software consolidating the original sets (DAS\_Tool, Binning\_refiner, metaWRAP). Only bins with ≥50% completion and ≤10% contamination are shown.
6. Figure S5: True recall and precision (determined with AMBER) of bins recovered from the CAMI’s high, medium, and low complexity synthetic data sets using original binning software (metaBAT2, MaxBin2, CONCOCT) and software consolidating the original sets (DAS\_Tool, Binning\_refiner, metaWRAP). Only bins with ≥0.5% recall and ≥0.9% precision are shown.
7. Figure S6: Completion of bins recovered from water, gut, and soil metagenomes with the metaWRAP-Bin\_refinement module with a varying minimum completion parameter (-c), but constant maximum contamination parameter (-x 10). The numbers in the brackets indicate the number of extra bins gained at that threshold compared to the baseline run (-c 50 –x 10). Only bins with ≥50% completion and ≤10% contamination are shown.
8. Figure S7: Contamination of bins recovered from water, gut, and soil metagenomes with the metaWRAP-Bin\_refinement module with a varying maximum contamination parameter (-x), but constant minimum completion parameter (-c 50). The numbers in the brackets indicate the number of extra bins gained at that threshold compared to the baseline run (-c 50 –x 10). Only bins with ≥50% completion and ≤10% contamination are shown.
9. Figure S8: Taxonomic distribution of reads from water, gut and soil metagenomes, estimated with the metaWRAP-Kraken module.
10. Figure S9: Clustered heat maps showing the log of bin abundance of bins extracted with metaWRAP-Bin\_refinement (-c 50 –x 10) across samples in water, gut, and soil metagenomes, calculated and plotted with metaWRAP’s Quant\_bins module.
11. Figure S10: Distribution of the taxonomy among Bacterial bins extracted from water, gut, and soil metagenomes using metaWRAP’s Bin\_refinement module (-c 50 –x 10). Taxonomy estimated with metaWRAP’s Classify\_bins module.
12. Figure S11: MetaWRAP-Blobology visualization of water, gut, and soil metagenomes, showing the GC and average coverage of each successfully binned contig (metaWRAP-Bin\_refinement –c 70 –x 10) in the assemblies, and annotated with the taxonomy at the phylum level, and the bins that they belong to (bin colors are chosen at random).

**Abbreviations**

WMG, whole metagenome; -c, minimum completion parameter; -x, maximum contamination parameter

**Acknowledgements**

We thank early users of metaWRAP for their patience and help with debugging.

**Funding**

This work was supported by grants NNX15AP18G and NNX15AK57G from NASA and grant DEB1556574 from the NSF and grant HG006620 from NIH/NHGRI

**Availability of data and materials**

MetaWRAP is an open source software available at GitHub <https://github.com/ursky/metaWRAP> and Anaconda <https://anaconda.org/ursky/metawrap-binning>. A step-by-step guide to using metaWRAP can be found at <https://github.com/ursky/metaWRAP/blob/master/Usage_tutorial.md>. All analysis results and scripts used to generate figures are available at https://github.com/ursky/metawrap\_paper.

Synthetic data used in benchmarking is from the original CAMI challenge https://data.cami-challenge.org/participate. Datasets used in this work are available at National Centre for Biotechnology Information under SRA numbers SRR2053273–SRR2053308 for the Central Baltic Surface Water Metagenome, SRA numbers ERR011087-ERR011136 for the Metagenomic of the Human Intestinal Tract (MetaHIT) survey, and at Joint Genome Institute under Gold Analysis Project IDs Ga0007435, Ga0007436, Ga0007437, Ga0007438, Ga0007439, and Ga0007440 for the soil data.

**Author's contributions**

GU created, released, and maintained the metaWRAP software, ran the benchmarks, and wrote the manuscript. JDR and JT provided ideas for improving metaWRAP and edited the manuscript. All authors read and approved the final manuscript.

**Ethics approval and consent to participate**

Not applicable

**Consent for publication**

Not applicable

**Competing interests**

The authors declare that they have no competing interests.

BIBLIOGRAPHY

1. Jovel J, Patterson J, Wang W, Hotte N, O'Keefe S, Mitchel T, Perry T, Kao D, Mason AL, Madsen KL *et al*: **Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics**. *Front Microbiol* 2016, **7**:459.

2. Mendes LW, Braga LPP, Navarrete AA, Souza DG, Silva GGZ, Tsai SM: **Using Metagenomics to Connect Microbial Community Biodiversity and Functions**. *Curr Issues Mol Biol* 2017, **24**:103-118.

3. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N: **Shotgun metagenomics, from sampling to analysis**. *Nat Biotechnol* 2017, **35**(9):833-844.

4. Wang WL, Xu SY, Ren ZG, Tao L, Jiang JW, Zheng SS: **Application of metagenomics in the human gut microbiome**. *World J Gastroenterol* 2015, **21**(3):803-814.

5. Guo J, Li J, Chen H, Bond PL, Yuan Z: **Metagenomic analysis reveals wastewater treatment plants as hotspots of antibiotic resistance genes and mobile genetic elements**. *Water Res* 2017, **123**:468-478.

6. Meyer KM, Klein AM, Rodrigues JL, Nusslein K, Tringe SG, Mirza BS, Tiedje JM, Bohannan BJ: **Conversion of Amazon rainforest to agriculture alters community traits of methane-cycling organisms**. *Mol Ecol* 2017, **26**(6):1547-1556.

7. Oulas A, Pavloudi C, Polymenakou P, Pavlopoulos GA, Papanikolaou N, Kotoulas G, Arvanitidis C, Iliopoulos I: **Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies**. *Bioinform Biol Insights* 2015, **9**:75-88.

8. Roumpeka DD, Wallace RJ, Escalettes F, Fotheringham I, Watson M: **A Review of Bioinformatics Tools for Bio-Prospecting from Metagenomic Sequence Data**. *Front Genet* 2017, **8**:23.

9. Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Droge J, Gregor I, Majda S, Fiedler J, Dahms E *et al*: **Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software**. *Nat Methods* 2017, **14**(11):1063-1071.

10. Piro VC, Matschkowski M, Renard BY: **MetaMeta: integrating metagenome analysis tools to improve taxonomic profiling**. *Microbiome* 2017, **5**(1):101.

11. Escobar-Zepeda A, Vera-Ponce de Leon A, Sanchez-Flores A: **The Road to Metagenomics: From Microbiology to DNA Sequencing Technologies and Bioinformatics**. *Front Genet* 2015, **6**:348.

12. Sharpton TJ: **An introduction to the analysis of shotgun metagenomic data**. *Front Plant Sci* 2014, **5**:209.

13. Ladoukakis E, Kolisis FN, Chatziioannou AA: **Integrative workflows for metagenomic analysis**. *Front Cell Dev Biol* 2014, **2**:70.

14. Batut B, Gravouil K, Defois C, Hiltemann S, Brugère J-F, Peyretaillade E, Peyret P: **ASaiM: a Galaxy-based framework to analyze raw shotgun data from microbiota**. *bioRxiv* 2017.

15. Kesh S, Raghupathi W: **Critical issues in bioinformatics and computing**. *Perspect Health Inf Manag* 2004, **1**:9.

16. Keegan KP, Glass EM, Meyer F: **MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function**. In: *Microbial Environmental Genomics (MEG).* Edited by Martin F, Uroz S. New York, NY: Springer New York; 2016: 207-233.

17. Chen IA, Markowitz VM, Chu K, Palaniappan K, Szeto E, Pillay M, Ratner A, Huang J, Andersen E, Huntemann M *et al*: **IMG/M: integrated genome and metagenome comparative data analysis system**. *Nucleic Acids Res* 2017, **45**(D1):D507-D516.

18. Louvel G, Der Sarkissian C, Hanghoj K, Orlando L: **metaBIT, an integrative and automated metagenomic pipeline for analysing microbial profiles from high-throughput sequencing shotgun data**. *Mol Ecol Resour* 2016, **16**(6):1415-1427.

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

20. Iyer R, Baliga NS, Camilli A: **Catabolite control protein A (CcpA) contributes to virulence and regulation of sugar metabolism in Streptococcus pneumoniae**. *J Bacteriol* 2005, **187**(24):8340-8349.

21. Wu YW, Simmons BA, Singer SW: **MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets**. *Bioinformatics* 2016, **32**(4):605-607.

22. Kang DD, Froula J, Egan R, Wang Z: **MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities**. *PeerJ* 2015, **3**:e1165.

23. Mande SS, Mohammed MH, Ghosh TS: **Classification of metagenomic sequences: methods and challenges**. *Brief Bioinform* 2012, **13**(6):669-681.

24. Imelfort M, Parks D, Woodcroft BJ, Dennis P, Hugenholtz P, Tyson GW: **GroopM: an automated tool for the recovery of population genomes from related metagenomes**. *PeerJ* 2014, **2**:e603.

25. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, Banfield JF: **Recovery of genomes from metagenomes via a dereplication, aggregation, and scoring strategy**. *bioRxiv* 2017.

26. Song WZ, Thomas T: **Binning\_refiner: improving genome bins through the combination of different binning programs**. *Bioinformatics* 2017, **33**(12):1873-1875.

27. Sharon I, Morowitz MJ, Thomas BC, Costello EK, Relman DA, Banfield JF: **Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization**. *Genome Res* 2013, **23**(1):111-120.

28. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, Darling A, Malfatti S, Swan BK, Gies EA *et al*: **Insights into the phylogeny and coding potential of microbial dark matter**. *Nature* 2013, **499**(7459):431-437.

29. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW: **CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes**. *Genome Res* 2015, **25**(7):1043-1055.

30. Wood DE, Salzberg SL: **Kraken: ultrafast metagenomic sequence classification using exact alignments**. *Genome Biol* 2014, **15**(3):R46.

31. Droge J, Gregor I, McHardy AC: **Taxator-tk: precise taxonomic assignment of metagenomes by fast approximation of evolutionary neighborhoods**. *Bioinformatics* 2015, **31**(6):817-824.

32. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features**. *Bioinformatics* 2010, **26**(6):841-842.

33. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.

34. Grüning B, Dale R, Sjödin A, Rowe J, Chapman BA, Tomkins-Tinch CH, Valieris R, Köster J: **Bioconda: A sustainable and comprehensive software distribution for the life sciences**. *bioRxiv* 2017.

35. Li D, Luo R, Liu CM, Leung CM, Ting HF, Sadakane K, Yamashita H, Lam TW: **MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices**. *Methods* 2016, **102**:3-11.

36. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA: **metaSPAdes: a new versatile metagenomic assembler**. *Genome Res* 2017, **27**(5):824-834.

37. Ondov BD, Bergman NH, Phillippy AM: **Interactive metagenomic visualization in a Web browser**. *BMC Bioinformatics* 2011, **12**:385.

38. Kumar S, Jones M, Koutsovoulos G, Clarke M, Blaxter M: **Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots**. *Frontiers in Genetics* 2013, **4**:237.

39. Meyer F, Hofmann P, Belmann P, Garrido-Oter R, Fritz A, Sczyrba A, McHardy AC: **AMBER: Assessment of Metagenome BinnERs**. *bioRxiv* 2017.

40. Hugerth LW, Larsson J, Alneberg J, Lindh MV, Legrand C, Pinhassi J, Andersson AF: **Metagenome-assembled genomes uncover a global brackish microbiome**. *Genome Biol* 2015, **16**:279.

41. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T *et al*: **A human gut microbial gene catalogue established by metagenomic sequencing**. *Nature* 2010, **464**(7285):59-65.

42. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD *et al*: **SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing.** *J Comput Biol* 2012, **19**:455-477.

43. Chen Y, Ye W, Zhang Y, Xu Y: **High speed BLASTN: an accelerated MegaBLAST search tool**. *Nucleic Acids Res* 2015, **43**(16):7762-7768.

44. Krueger F: **Trim Galore!: a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files**. In*.*, 0.4.5 edn: Bioconda; 2015.

45. Agarwala R, Morgulis A: **BMTagger aka Best Match Tagger is for removing human reads from metagenomics datasets**. In*.*, vol. 3.101: Bioconda; 2010.

46. Brown J, Pirrung M, McCue LA: **FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool**. *Bioinformatics* 2017.

47. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.

48. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nat Methods* 2012, **9**(4):357-359.

49. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C: **Salmon provides fast and bias-aware quantification of transcript expression**. *Nat Methods* 2017, **14**(4):417-419.

50. **DIBSI Metagenomics Workshop at UC Davis** [<http://2017-dibsi-metagenomics.readthedocs.io/en/latest/>]

51. Seemann T: **Prokka: rapid prokaryotic genome annotation**. *Bioinformatics* 2014, **30**(14):2068-2069.