**METAWRAP IS AN EASY-TO-USE 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). With the help of Anaconda, metaWRAP is easy to download and install for biologists without significant computation experience. The metaWRAP installation produces a bioinformatics environment with over 150 commonly used bioinformatics software and libraries, saving the user from installing and configuring them individually (Figure S1). Each of metaWRAP’s modules is a standalone program, allowing the user to choose only the functions they are interested in. Alternatively, the user may follow the intuitive workflow starting from raw metagenomic sequencing reads (Figure 1).

|  |
| --- |
| Figure 1. Overall workflow of using the modules of metaWRAP (in red). Arrows represent major input/output points of modules, but not all required inputs are shown. |

The metaWRAP workflow starts with modules that conveniently wrap common software to quality-control and taxonomically profile the reads, assemble them, and then bin the assembly. MetaWRAP::Read\_qc module trims the raw sequence reads and removes human contamination from each of the sequenced samples. Quality control reports are also generated to evaluate the sequencing quality. The reads from all given samples are then assembled with the metaWRAP::Assembly module, also producing an assembly report. Both the reads from each sample and the assembly can be taxonomically profiled with the Kraken module, producing interactive kronagrams of community taxonomy. The metaWRAP::Binning module is then used to bin the scaffolds of the joint assembly using three metagenomic binners – MaxBin2, metaBAT2, and CONCOCT.

The rest of metaWRAP’s modules focus on refining and analyzing metagenomic bins. It is important to note that the user can provide their own sets of bins produced with software besides the ones available in metaWRAP::Binning. These bin sets are then passed to the metaWRAP::Bin\_refinement module, which hybridizes the bin sets with Binning\_refiner, and then finds the best version of each bin based on completion and contamination metrics estimated with CheckM (Figure S2). The scaffolds in the final bin set is then de-replicated, and a report of their completion, contamination, and other metrics is produced. These bins can then be visualized by using the metaWRAP::Blobology module, which plots the contigs of the joint assembly on a GC vs abundance plot, and 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 with Salmon. MetaWRAP::Classify\_bins can be used to conservatively, but accurately estimate their taxonomy with the use of Taxator-tk. Finally, metaWRAP::Reassemble\_bins can be used to reassemble the reads belonging to each bin, improving their N50, completion, and contamination (Figure S3).

**METAWRAP::BIN\_REFINEMENT IMPROVES BIN PREDICTION IN CAMI SYNTHETIC DATA**

In order to test the efficacy of the metaWRAP::Bin\_refinement module in improving bins based on three different bin sets, we tested applied the module to the data set from the CAMI challenge, which included synthetically generated metagenomes with high, medium, and low diversity. The “golden standard” assembly from each CAMI challenge was binned with metaBAT2 v2.12.1, Maxbin2 v2.2.4, and CONCOCT v0.4.0 using the metaWRAP::Binning module. The resulting three bin sets were then refined with DAS\_Tool, binning\_refiner, and metaWRAP::Bin\_refinement 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, 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, which compared the bins against the true original genomes (Figure S4). Bin recall and precision were converted to completion and contamination percentages for easier viewing (Figure 2).

|  |
| --- |
| Figure 2. Completion and contamination of bins recovered from the CAMI binning challenge synthetic data sets using different binning strategies. The bin sets in dashed lines (metaBAT2, MaxBin2, CONCOCT) are original sets, while the bin sets in solid lines (DAS\_Tool, Binning\_refiner, metaWRAP) are bins produced by combining the original three sets. Only bins with a recall of greater than 50% and completion and less than 10% contamination are shown. |

Between the original binning software, 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. Finally, CONCOCT performed rather poorly in all but the smallest CAMI challenge data sets, producing 58 high quality bins and 40 near-perfect bins.

DAS\_Tool consistently produced high-completion bins in all CAMI challenges. However, these bins had relatively high contamination when compared to other bin refiners - Binning\_refiner and metaWRAP:: Bin\_refinement. Between all three challenges, DAS\_Tool was able to produce 426 high quality bins and 263 near-perfect bins. Binning\_refiner produced very pure bins, but did so at the expense of significantly reduced completion. In total, Binning\_refiner was able to produce 289 high quality bins and 210 near-perfect bins.

MetaWRAP::Bin\_refinent was able to consistently produce high completion and low contamination bins, coming close to DAS\_Tool in completion, and coming close to Binning\_refiner in contamination. In total, metaWRAP produced 457 high quality bins and 339 near-perfect bins. (Figures 2, S4)

**REAL BENCHMARKING DATA SETS**

MetaWRAP was also benchmarked against real metagenomic data. To test its efficacy on a range of microbial communities, it was tested on a water, gut, and soil microbiome WGS metagenomic data sets. The water data sets came from a brackish water survey of the Baltic Sea (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4699468/)>, and includes 36 samples (SRR2053273–SRR2053308) for a total of 196Gbp in sequencing data. The gut data set came from the Metagenomic of the Human Intestinal Tract (MetaHIT) survey (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3779803/)>, and consisted of 50 samples (ERR011087- ERR011136) and a total of 144Gbp. Finally, the soil data set came from grassland soil microbial communities from Angelo Coastal Reserve (<https://www.biorxiv.org/content/early/2017/02/11/107789.full.pdf+html)>, and included 6 samples (Gold Analysis Project IDs: Ga0007435, Ga0007436, Ga0007437, Ga0007438, Ga0007439, Ga0007440) and 481Gbp of sequencing data.

The samples from each microbiome type were run through metaWRAP’s Read\_qc module to trim the reads and remove human contamination, and then the Kraken module of metaWRAP was run on the reads to investigate 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 were comprised of a wide variety of Proteobacteria and Terrabacteria.

**METAWRAP::BIN\_REFINEMENT IMPROVES BIN PREDICTION IN REAL DATA SETS**

The quality-controlled reads were then co-assembled with MegaHit by using the Assembly module. Contigs shorter than 1000bp were discarded, with the exception of the soil assembly, for which the cutoff of 3000bp was chosen to shorten the binning time. The contigs were then binned with metaBAT2 v2.12.1, Maxbin2 v2.2.4, and CONCOCT v0.4.0 using the metaWRAP::Binning module. The resulting three bin sets of each microbiome type were then passed to DAS\_Tool, Binning\_refiner, and metaWRAP::Bin\_refinement to attempt to improve the bin sets. MetaWRAP was run with a variety of minimum completion and maximum contamination settings (-c and -x) depending on the bin quality of interest. To benchmark the bins produced by all the binning methods, the completion and contamination of the bins was estimated with CheckM.

To compare the full range of bins metaWRAP is capable of producing, metaWRAP:bin\_refinemnt was run with –c 50 –x 10 setting and the bins compared to those produced by other binning methods. Bins with a completion greater than 50% and a contamination less than 10% were sorted by completion and contamination (Figure 3).

|  |
| --- |
| Figure 3. Completion and contamination of bins recovered from real metagenomic data sets by using different binning strategies. The bin sets in dashed lines (metaBAT2, MaxBin2, CONCOCT) are original sets, while the bin sets in solid lines (DAS\_Tool, Binning\_refiner, metaWRAP) are bins produced by combining the original three sets. Only bins with a completion greater than 50% and contamination less than 10% are shown (as estimated by CheckM). |

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 came out with 151, 98, and 40 bins, and CONCOCT with 65, 121, and 39 bins (Figure 3).

Despite incorporating all the binning methods, DAS\_Tool struggled to improve the metaBAT2 bins in all three data sets, producing 198, 130, and 63 acceptable quality bins in water, gut, and soil samples, respectively. However, DAS\_Tool came out ahead at higher bin completion ranges (>80%), although at the expense of increased contamination. Binning\_refiner also produced a similar number of bins in the acceptable quality range when compared to metaBAT2, with 206, 138, and 83 bins in water, gut, and soil data sets. The resulting bins were less complete in the gut data sets, especially at the higher completion ranges. However, Binning\_refiner was able to significantly reduce the contamination of bins in all the data sets when compared to to the original bin sets. Finally, metaWRAP::Bin\_refinement was able to significantly improve the original bins of all three data sets, producing 235, 175, and 134 acceptable quality bins in water, gut, and soil samples. MetaWRAP was also able to match DAS\_Tool’s high completion and Binning\_refiner’s low contamination.

**PARAMETERIZATION OF METAWRAP::BIN\_REFINEMENT**

To demonstrate the effects of changing the –c (minimum completion) and –x (maximum contamination) parameters of metaWRAP’s Bin\_refinement module, metaWRAP was re-run on the original bin sets with varying minimum completion (but fixed maximum contamination), and varying maximum contamination (but fixed minimum completion). The number of extra bins that was gained at that quality threshold when compared to the original run (-c 50 –x 10) was noted.

When changing the minimum completion parameter of metaWRAP but keeping the maximum contamination at 10%, the number of bins recovered at each threshold was notably higher when compared to the original metaWRAP run (-c 50 –x 10), although the improvements in the soil bins were not as significant (Figure S5). The improvements were especially noticeable at higher completion ranges. MetaWRAP –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 metaWRAP –c 50 –x 10.

When changing the maximum contamination parameter of metaWRAP, but keeping the minimum completion at 50%, the number of bins recovered at each threshold was higher when compared to the original metaWRAP run (-c 50 –x 10) (Figure S6). Just as with varying -c, the improvements in the soil bins were not as significant. These improvements were especialy significant at the lower contamination range. MetaWRAP with –c 50 –x 1 parameters extracted 8, 21, 4 (water, gut, and soil, respectively) more bins at a maximum contamination of 1%, when compared to metaWRAP –c 50 –x 10.

**METAWRAP::REASSEMBLE\_BINS SIGNIFFICANTLY IMPROCED BIN QUALITY**

The bins produced from metaWRAP::Bin\_refinement module run (–c 50 –x 10) were also run through the metaWRAP::Reassemble\_bins module, and resulting bins were re-evaluated with CheckM. Of the 235 bins from the water data set, 139 were improved through strict reassembly, 45 though permissive reassembly, and 51 remained unchanged. Of the 175 bins from the gut data set, 82 were improved through strict reassembly, 90 though permissive reassembly, and 3 remained unchanged. Finally, of the 134 bins from the soil dataset, only 3 were improved through strict reassembly, and the rest remained unchanged. (Figure 4)

|  |
| --- |
| Figure 4. Comparison of N50, completion, and contamination metrics of original bins and bins reassembled with metaWRAP’s Reassemble\_bins module, as evaluated by CheckM. Only bins with a completion greater than 50% and contamination less than 10% are shown. |

With the exception of the soil bins, the bin sets were significantly improved by the Reassemble\_bins module. In water bins, the N50 increased by an average of 121% +/- 151.5%, the completion increased by an average of 1.49% +/- 2.41%, and the contamination reduced by an average of 0.85% +/- 1.13%. In the gut bins, the N50 increased by an average of 128.8% +/- 85.3%, the completion increased by an average of 2.48% +/- 2.89%, and the contamination reduced by an average of 1.21% +/- 1.13%. Because only 3 out of the 134 soil bins were successfully reassembled, the improvements to the set as a whole were minor.

**METAWRAP PRODUCES THE MOST HIGH-QUALITY BINS**

We investigated the performance of different binning approaches when extracting high quality putative genomes, with a contamination less than 5% and completion greater than 70%, 80%, 90%, and 95%. The number of bins meeting these criteria were counted in the original bin sets (metaBAT2, MaxBin2, CONCOCT) and binning refining software (metaWRAP, Binning\_refiner, DAS\_Tool). MetaWRAP was run with a maximum contamination of 5% and a minimum completion of 70% (metaWRAP’s default), 80%, 90%, or 95%. (Figure 5)

|  |
| --- |
| Figure 5. Number of high purity bins (<5% contamination) extracted with 70%, 80%, 90%, and 95% completion using original binning software (metaBAT2, MaxBin2, and CONCOCT) and bin refining algorithms (Binning\_refiner, DAS\_Tool, metaWRAP, and metaWRAP with reassembly). Binners labeled mW and mW.R represent metaWRAP::Bin\_refinement and metaWRAP::Reassemble\_bins runs with varying minimum completion parameters. The reassembles were done on metaWRAP::Bin\_refinment iterations with a minimum completion of 60%, 70%, 80%, and 90%, respectively. Completion and contamination were estimated with CheckM. |

MetaBAT2 performed the best out of the three original binning software, producing the most high-quality bins at every completion cut off, with the exception of gut bins of a minimum completion of 95%. The default run of metaWRAP consistently produced the highest number of high-quality putative genomes at every completion cut off and in every sample type, with the exception of gut bins at 95% completion. The default metaWRAP produced 185, 158, 117, and 64 bins (at 70%, 80%, 90%, and 95% completion, respectively) from the water data set, 132, 103, 66, and 33 bins from the gut data set, and 47, 29, 15, and 6 bins from the soil data set. These numbers further improved when re-running metaWRAP with minimum completion setting corresponding to the bin quality range being considered. Custom metaWRAP runs returned 185, 164, 128, and 72 water bins, 132, 107, 76, and 46 gut bins, and 47, 30, 15, and 6 soil bins. When providing metaWRAP::Bin\_refinement module the target bin quality range, it outperformed every other tested binning and bin refinement method at every quality threshold.

The reassembly of the bins with metaWRAP::Reassemble\_bins made a significant improvement on the number of high-quality putative genomes extracted from the gut and water data sets. First, bin sets were produced with metaWRAP’s Bin\_refinement module with a maximum contamination setting of 10% and minimum completion settings of 60%, 70%, 80%, and 90%. These bin sets were then reassembled with the Reassemble\_bins module with a maximum contamination setting of 5% and minimum completion settings of 70% (default setting), 80%, 90%, and 95%, respectively. Default reassembly produced 199, 174, 132, and 75 water bins and 146, 120, 78, and 42 gut bins with 70%, 80%, 90%, and 95% completion, respectively. This was a significant improvement compared to 185, 158, 117, and 64 water bins and 132, 103, 66, and 33 gut bins from the non-reassembled sets. When parameterizing the metaWRAP::Reassemble\_bins module based on the threshold being considered, the module produced 199, 173, 134, and 76 water bins and 146, 119, 80, and 53 gut bins. This is an even further improvement compared to 185, 164, 128, and 72 water bins and 132, 107, 76, and 46 gut bins produced from non-reassembled parameterized Bin\_refinement runs. The number of bins from produced from the soil dataset remained unchanged.

**METAWRAP OFFERS ANALYSIS AND VISUALIZATION OF METAGENOMIC BINS**

The bins produced by the metaWRAP::Bin\_refinemnt module (comp>50%, cont<10%) were then run through the metaWRAP::Classify\_bins module. The bins from the data sets belonged to a wide range of phyla, but most of the water bins were Proteobacteria and Bacteroidetes, most of the gut bins were Firmicutes, and most of the soil bins were Proteobacteria (Figure S10). The soil bins were also the most difficult to assign taxonomy to, with only 37% of the bins classified to the Phylum level (Figure S9).

The metaWRAP::Quant\_bins module was used to estimate the abundance of each bin across samples, and the results are shown in a heatmap (Figure S8). Finally, the metaWRAP::Blobology module was used to visualize the taxonomic composition of the water, gut, and soil communities with the use of Taxon-Annotated-GC-Coverage plots (TAGC plots). The module was also used to annotate the bins (comp>70%, cont<10%) that the contigs belonged to, allowing for visual inspection of the binning process success (Figure 6). Finally, the successfully binned contigs were also annotated with BLAST-determined taxonomy (Figure S11).