The study of microbial communities through whole genome metagenome (WMG) sequencing opens new opportunities for the study of not only the taxonomic composition of microbiomes, but also their metabolic potential. However, because all the sequencing reads from different community members are mixed together, this poses a unique and difficult challenge for analysis and interpretation of such sequencing data. With rapid improvements in sequencing technologies, the number of sequenced metagenomes is rapidly growing, as is the need to more sophisticated analysis software.

There are now thousands of available software to analyze metagenomic data, and this number is rapidly growing. In order to perform meaningful analysis of raw metagenomic data, microbial biologists have to use dozens of software and in-house scripts to take the raw sequencing reads through quality trimming, de-contamination, assembly, taxonomic profiling, binning, functional annotation, and many other functions that they may be interested in. The establishment of a such a pipeline is a difficult task for a variety of reasons. First, there are dozens of competing software that can perform each step, and the investigator needs to research the best currently available tool for their task. Next, the software needs to be installed, which is a major challenge faced not only by biologists without significant computational experience, but even by bioinformaticians. Also, a sophisticated metagenomics analysis pipeline can include hundreds of software and libraries, resulting in computation environment conflicts that are difficult to resolve. The vast majority of computational analysis requires significant resources, which means it needs to be performed on Unix clusters, further deterring biologists. Finally, running each software and script one after another is a lot of work. Each output needs to be converted into the correct format to the input for the next step, and errors need to be detected, identified, and handled.

Together, these challenges present a major burden on anyone attempting metagenomic analysis. For trained and experienced bioinformaticians, such pipelines take a lot of time and effort to establish and run. For microbial biologists without significant computational experience however, it is incredibly difficult and intimidating to even begin such an analysis. This prevents many microbial biologists from attempting such analysis on their own, forcing them to outsource bioinformatics analysis or perform only surface-level inspection of the data. These barriers preventing biologists from properly utilizing the wealth of available computational power significantly slows the progress of microbiology as a field. With this in mind, metaWRAP was created in order to make metagenomic analysis easily accessible.

One aspect of metagenomic analysis is de-convoluting a WMG by extracting the single genomes of its community members. This process is called binning, and is typically accomplished by assembling the WMG reads into a metagenomic assembly, and then sorting the resulting scaffolds based on predicted genome membership. Genome-resolved metagenomics allows for inspection of the metabolic pathways and abundance shifts of individual organisms, increasing the resolution of metagenomic analysis.

Metagenomic binning poses a great computational challenge, as the goal is to predict scaffolds belonging to the same organisms based on abundance and sequence qualities alone. Many powerful tools have been developed to tackle this problem, each one using a different approach to make the predictions.

Many metagenomic binning tools rely on k-mer composition and codon usage properties of scaffolds, and use emergent self-organizing maps (ESOMs) to make bin predictions. A major assumption that most binners also make is that quality that scaffolds belonging to the same organism are expected to have similar read coverage in any sample. This concept greatly improves bin prediction by clustering together scaffolds that have similar abundances across multiple samples. Finally, as the number of reference microbial sequences grows, some tools also incorporate the phylogenetic information that can be derived from sequences and their genes.

Another challenge of metagenomic bin prediction is evaluating the quality of the resulting bins. In order for a metagenomic bin to be considered the genome of a single organism, it must cover a significant length of the true genome (have a high completion), and also not have sequences belonging to other organism (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. A perfect bin is expected to have all of such genes, but if some of them come in multiple copies, it is assumed that they come from contamination from other organisms. Assuming a random distribution of universal single-copy genes across a genome, the completion of a bin is roughly proportional to the percentage of expected universal single-copy genes that it has. Similarly, the contamination of a bin can be predicted by the percentage of universal single-copy genes that are found in multiple copies. CheckM is a powerful tool that takes this idea to the next level, inspecting not only the universal single-copy genes of all domains of life, but the single-copy genes that a genomes of a particular taxonomy is expected to have.

Metagenomic binning tools are typically evaluated and benchmarked by their creators on relatively simple synthetic and real data sets, demonstrating their superiority in the specific data sets. To more objectively compare the performance of different approaches, a various binning software were challenged to bin a complex synthetic data set in the Critical Assessment of Metagenomic Interpretation (CAMI) study. While this offered a lot of insight into how binning software compare, it also made clear that there is no one software that performs best all the time. In addition, some genomes in the same sample were better recovered by some binners that others.

To combine the strengths and minimize weaknesses of different binning software, a couple of packages were created to consolidate the predictions from multiple binning software. DAS\_Tool does this by aggregating, de-replicating, and collapsing bins from multiple bin sets. The software predicts single-copy genes in all the provided bin, aggregated bins with overlapping genes, and extracts a more complete consensus bin from each aggregate. This aggressive bin collapsing approach significantly improves the completion of the bin predictions with minimal increase of contamination. Binning\_refiner takes the opposite approach – it splits the contigs into bins such that all the contig division boundaries of the original bin predictions are satisfied. This breaks up the contigs into many more bins, reducing their completion, but it also greatly reduces their contamination. While taking opposite approaches consolidate sets of bins from different software and result in a superior bin set, they have limitations – DAS\_Tool focuses on completion at expense of introducing contamination, while Binning\_refiner prioritizes purity, but loses completeness. This problem inspired the creation of the meraWRAP::Bin\_refinement module, which is able to reduce contamination and improve completion of bin sets by having both a splitting and a collapsing step.

Another way to potentially improve bin quality is bin reassembly – extracting reads that belong to a given bin and assembling them separately from the rest of the metagenome. This feature was featured in MaxBin, but it was highly experimental and was not benchmarked. Bin reassembly has potential to improve sequence contiguity by closing gaps between contigs and increase bin completion by recruiting read pairs that only had one read incorporated into the original assembly. However, there is also a risk of introducing contamination from other organisms, especially when there are multiple closely related community members that have a high sequence homology. Still, with proper benchmarking and reassembly evaluation this approach has promise to significantly improve at least some bins in a microbial community, therefore result in a more complete function annotation. This idea lead to the creation of the metaWRAP::Reassemble\_bins module.

While it is a challenge to extract high-quality bins, there is also a lack of software to inspect, analyze, and visualize them. While there are many software that can accurately predict the taxonomy of metagenomic scaffolds (such as Taxator-tk), there is no software to classify entire metagenomic bins. Similarly, there are many ways to estimate the coverage of scaffolds based on read alignment depth, but no software that can find the coverages of entire bins across many samples. Additionally, kmer based transcript quantitation algorithms like Salmon, which were originally designed to quantify transcript expression in RNAseq, show promise in improving depth estimates in metagenomic scaffolds. These knowledge gaps inspired the creation of metaWRAP’s Quant\_bins and Classify\_bins modules.

Visualizing metagenomic communities with Taxon-Annotated-GC-Coverage plots (or blob plots) with software such as Blobology offers a way to compare and understand general community composition and structure. However, these blobplots can also be annotated with bin membership, helping visualize binning success.