# Assembly pipeline overview

All MAG and SGB scripts can be found [here](https://www.dropbox.com/scl/fo/ajy70v3hmeuzv48pnku30/h?dl=0&rlkey=6xqvr7jwr5fynvr21eugfgaa3).

1. (If running biobakery workflows) Deconcatenation from kneaddata output
   1. Split into pair\_1, pair\_2, unmatched\_1, unmatched\_2
2. [MEGAHIT](https://www.metagenomics.wiki/tools/assembly/megahit) with pair\_1 and pair\_2 as paired ends and unmatched\_1 and unmatched\_2 as unpaired ends and --min-contig-len=1500
   1. Get final.contigs.fa file
3. Build a [bowtie2](http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml) index from the contigs
4. Align the reads to the bowtie2 index with --very-sensitive-local and --no-unal
5. Sort the bam files
6. Count alignments with [jgi\_summarize\_bam\_contig\_depths](https://bitbucket.org/berkeleylab/metabat/src/master/)
   1. Get contig\_depths.txt file
7. Run [MetaBAT](https://bitbucket.org/berkeleylab/metabat/src/master/) with --unbinned -m 1500 -t 3 (minimum contig length 1500, keep unbinned contigs in a file)
   1. Get folder of bins
8. Run [checkm](https://github.com/Ecogenomics/CheckM/wiki) coverage to get read coverage of bins
   1. Get coverage.tsv
9. Run checkm profile to get abundances of each bin (normalized by genome size)
   1. Get abundance.tsv
10. Use [samtools](http://www.htslib.org/doc/samtools-view.html) to count the number of reads that aligned, count the number of sequences in the input files
    1. Get mapped\_read\_num.txt
11. Calculate the n50 with mag\_n50\_calc.py and [assembly-stats](https://github.com/sanger-pathogens/assembly-stats)
    1. Get mags\_n50.tsv
12. Run checkm2 to get completeness and contamination for each bin
    1. Get checkm\_qa.tsv
13. Run checkm\_wrangling.R to merge in the n50 values
    1. Get checkm\_qa\_combined.tsv
14. Run [phylophlan\_metagenomic](https://github.com/biobakery/phylophlan/wiki) with --add\_ggb --add\_fgb to assign taxonomy
    1. Get combined.tsv
15. Run phylophlan\_tax\_wrangling.R to keep only taxonomies assigned at <0.05 Mash for species, <0.15 for genus, <0.3 for family
    1. Get phylophlan\_tax\_merged.tsv
16. Run mash\_list\_inputs.py to identify medium and high quality bins not assigned by phylophlan
17. Calculate the [Mash distance](https://github.com/marbl/mash) between every pair of these bins, place the bins in species genome bins (SGBs) based on <5% Mash distance, find the ANI between SGBs and merge the SGBs if the ANI is >95%
18. Merge these into the phylophlan output as new entries with names like sgb\_1
19. Run merge\_tax\_and\_abundance.R to take all the abundance.tsv files and all the mapped\_read\_num.txt files and put those abundances towards the relevant bins for each sample.