Overview of generic* metagenomics workflow

*This is generic; specific workflows can vary on the order of steps here and how they are done.

When working with your own data you should never follow any pipeline blindly. There can be critical differences based on your data.

sequencing facility → **fastq files**

@HISEQ2500:282:1:1101:1220:1944 1
ATCGGATCG...
+
<G.<G<AGGII...

might be done by sequencing facility

demultiplex (split samples by barodes)

Some tools:
- sabre
- fastx_demux (usearch/vsearch)
- idemp
- fastx barcode splitter (fastx-toolkit)

fastqc/multiqc

quality filter/trim (remove adapters/**primers**)

Some tools:
- trimmomatic
- bbduk.sh (bbtools suite of tools)

fasta/q files

>HISEQ2500:282:1:1101:1220:1944 1
ATCGGATCG...

no-assembly path

read-based analysis

Some tools:
- TIPP/SEPP
- metaphlan2
- humann2
- sourmash
- kraken

Count Table

	Sample_A	Sample_B	. . .
obj_1	0	428	. . .
obj_2	306	323	. . .
obj_3	217	1	. . .
. . .	. . .	. . .	. . .

Analysis

Some tools:
- phyloseq
- Breakaway
- DivNet
- CORNCOB
- SpiecEasi
- MaAsLin
- DESeq2

assembly path

consider testing assemblies with and w/o

digital normalization

Some tools:
- bbnorm
- diginorm

MetaQUAST is a great tool for comparing assemblies

(co)-assembly

Some assemblers and tools:
- Megahit (assembler)
- SPAdes (assembler)
- idba-ud (assembler)
- MetAMOS (assembler and analysis pipeline)
- MetaCompass (reference-guided)
- MetagenomeScope (visualize assembly graphs)

map individual sample reads to (co)-assembly

Generate coverage information (mapping)

Some tools:
- bowtie2
- bwa

Gene calling Functional/taxonomic profiling

Some tools:
- prodigal (identifies open reading frames)
- prokka (runs prodigal and performs annotations)
- GHOSTKOALA (web-hosted KEGG annotations)
- BLAST (protein nr db/refseq/COGs)

Recovering genomes from metagenomes

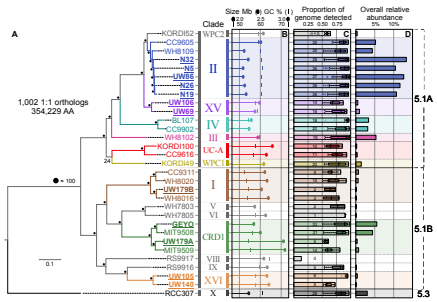
A note on MAGs:
MAGs (metagenome-assembled genomes) are not the same thing as isolate genomes. They are composite representative genomes of closely related genomic lineages.

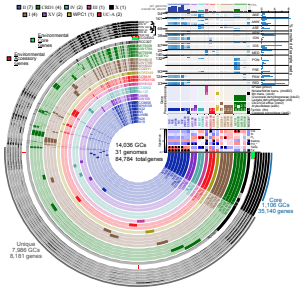
Some tools:
- anvi'o (interactive manual curation of bins; and much more)
- CONCOCT (kmer-based and coverage-based binning; also incorporated in anvi'o)
- COCACOLA (kmer-based, coverage-based, and incorporates paired-read linkage of contigs)
- MetaBAT2 (kmer-based and coverage-based binning tool)
- BinSanity (primarily coverage-based, optional second round kmer-based binning tool)
- checkm (genome-level taxonomy; and much more)
- DASTool (a tool for evaluating bins recovered by different methods)
- DESMAN (tool aimed at resolving strains)

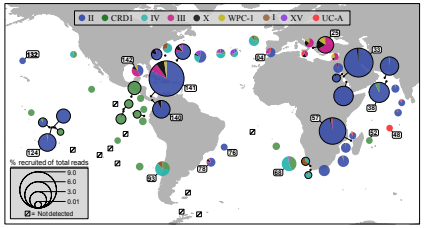
Some common genomics stuff

Phylogenomics
Comparative genomics
Pangenomics
Env. distributions

Some tools:
- anvi'o (integrated HMMs for common single-copy gene sets; integrated pangenomic workflow for identifying orthologs via OrthoMCL)
- PanOCT (identifies orthologs utilizing synteny information)
- StrainPhlAn/PanPhlAn (tools for strain-level analyses)
- MUSCLE (alignment software)
- FastTree (very fast, pseudo-maximum likelihood tree builder)
- RAxML (maximum likelihood tree builder)
- Mauve (whole-genome alignment)







astrobiomike.github.io