Assembly: annotation

Assembly validation & QC

PE metagenomic reads

QC (fastQC)

Trim adapters & quality filter (bbduk.sh)

Filtered reads (fastq)

Assemble (megahit)

Contigs (fasta)

Collect stats (stats.sh, MetaQUAST)

N50, max, min, mean

Map reads (bbmap.sh)

SAM files

Collect stats (bbmap.sh, samtools, piccard tools)

% reads mapped contig coverage

Gene prediction (Prodigal)

Amino Acid Seq (fasta)

Gene annotation (GenDB, MG-RAST, PFam, KEGG, COG)

General feature format (GFF)

PROKKA (prediction & annotation)

If we extracted abundance information from this output, what would it tell us?

GFF file
1. seqname
2. source
3. feature
4. start
5. end
6. score
7. strand
8. frame
9. group

Assemblies are not quantitative