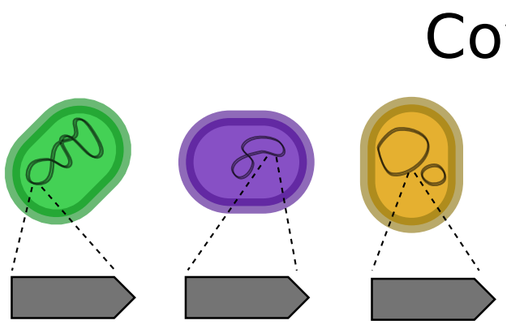
WRF's tips on environmental DNA bioinformatics

✓ = a reasonable way to plot the data

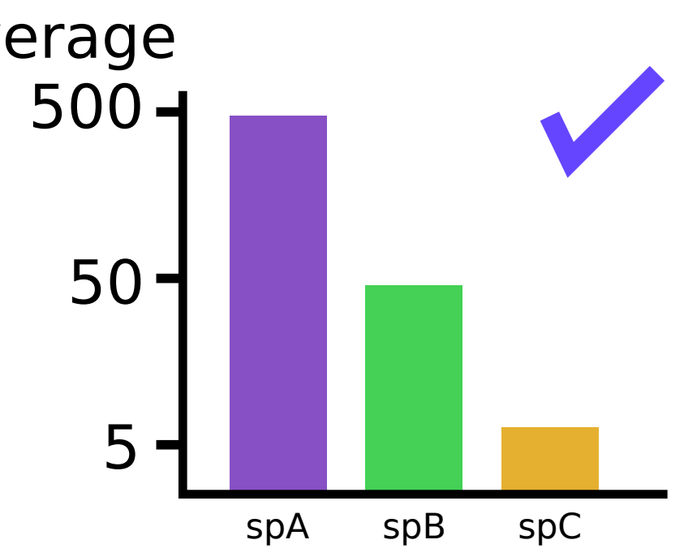
✗ = graph may be confusing or misleading

Amplicon

This refers to amplifying a single locus, often ribosomal RNA, across all species within a sample.
This typically informs who is in a sample, but not what they are doing.

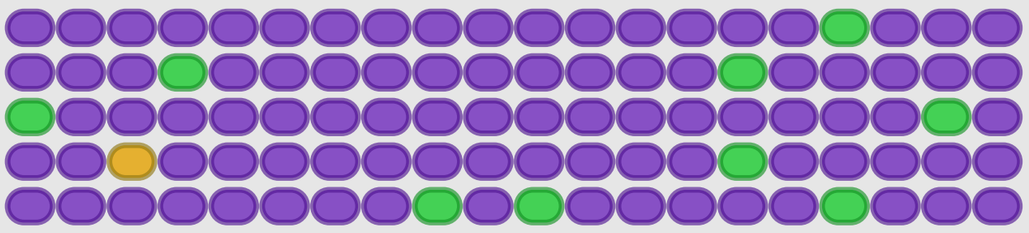


same single gene from multiple species



✓

Coverage means ***relative*** abundance in that sample

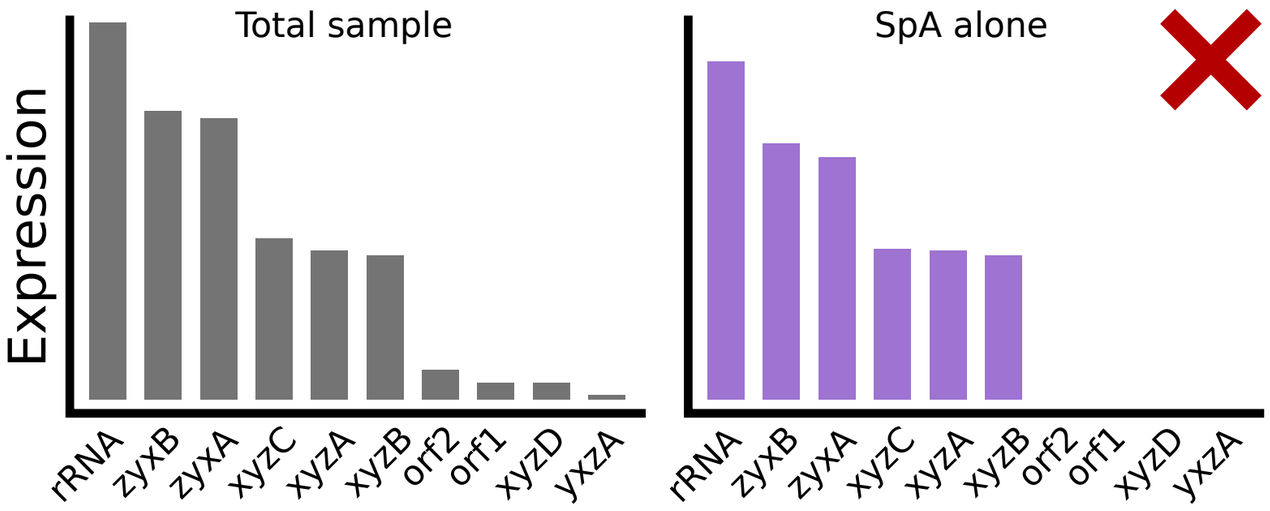


Metatranscriptomics

Ultimately, to understand the processes that occur in the sample, transcriptomes (snapshot of genes used at that time) are needed.

By itself, a transcriptome may indicate some dominant processes, i.e. certain genes or pathways that are highly expressed.

Transcript expression level and species abundance are both varied, so this profile may be dominated by one species in the sample.

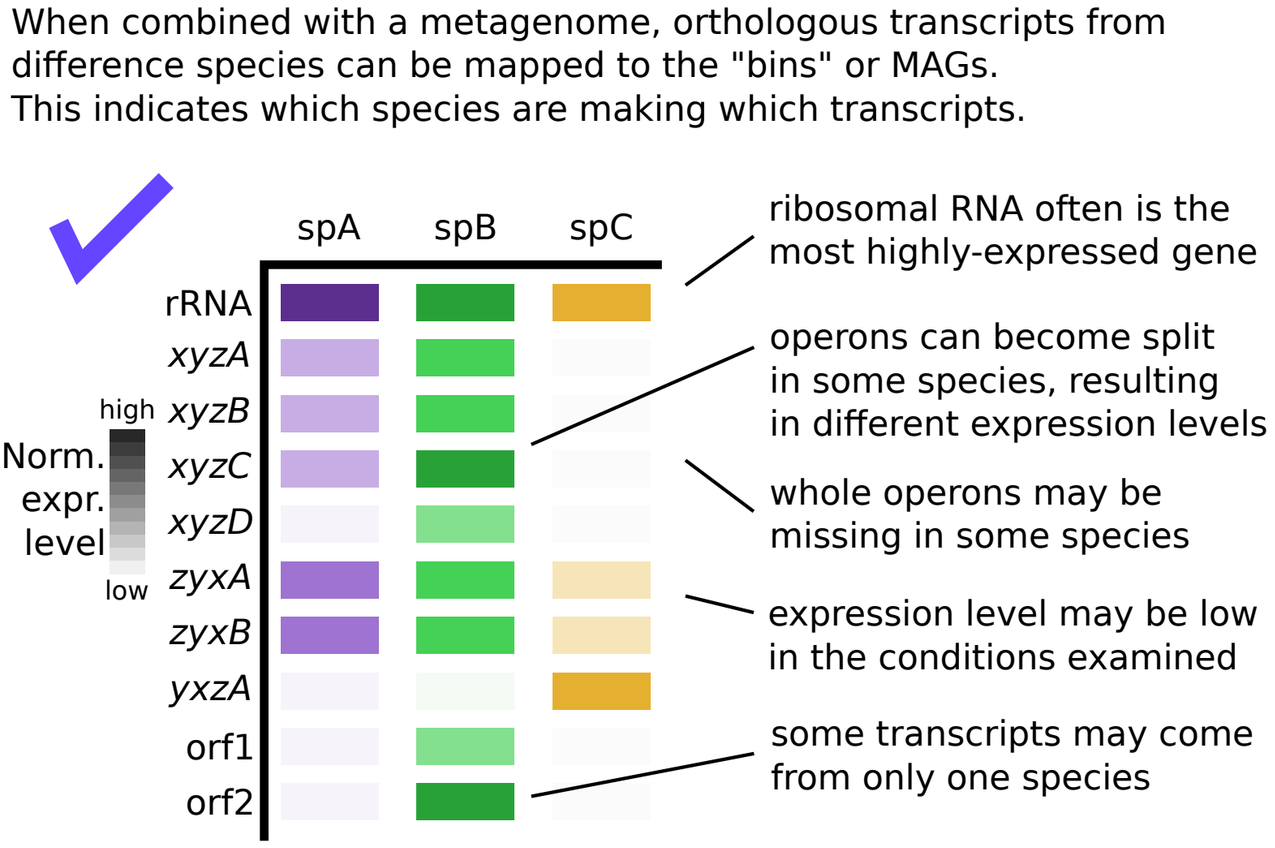


✗

However, without a metagenome, it is not clear to which species any given transcript may belong.

When combined with a metagenome, orthologous transcripts from difference species can be mapped to the "bins" or MAGs.
This indicates which species are making which transcripts.

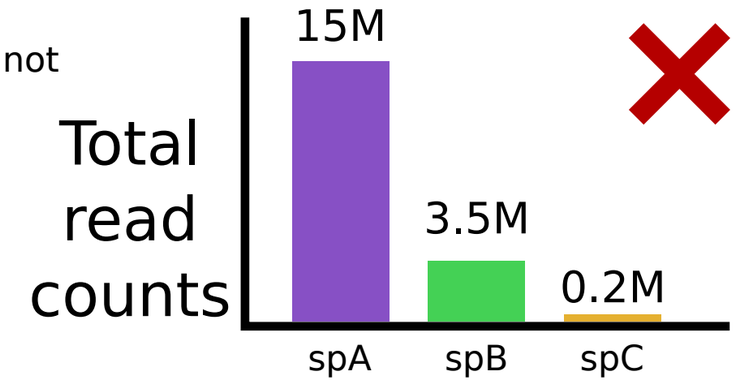
✓



Metagenomics

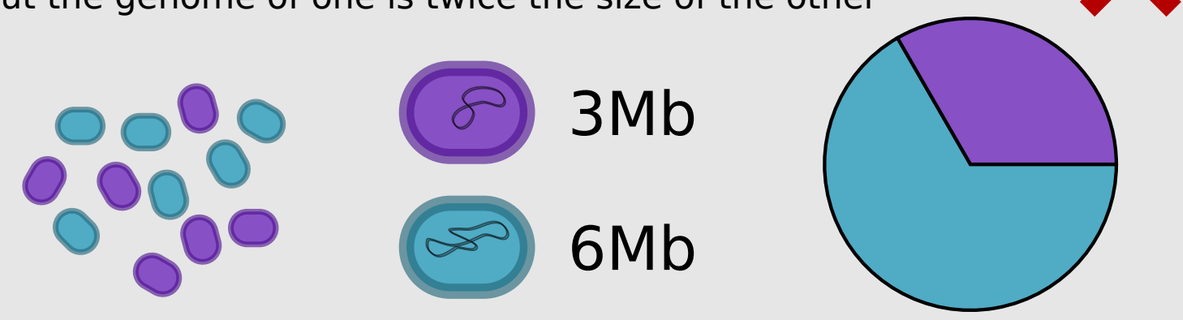
If all DNA in a sample is sequenced, this can give information about the composition of species, but also the genes of those species, i.e. what pathways or genes are potentially used.

Unassembled read counts are not representative of abundance, as they reflect both coverage and genome size



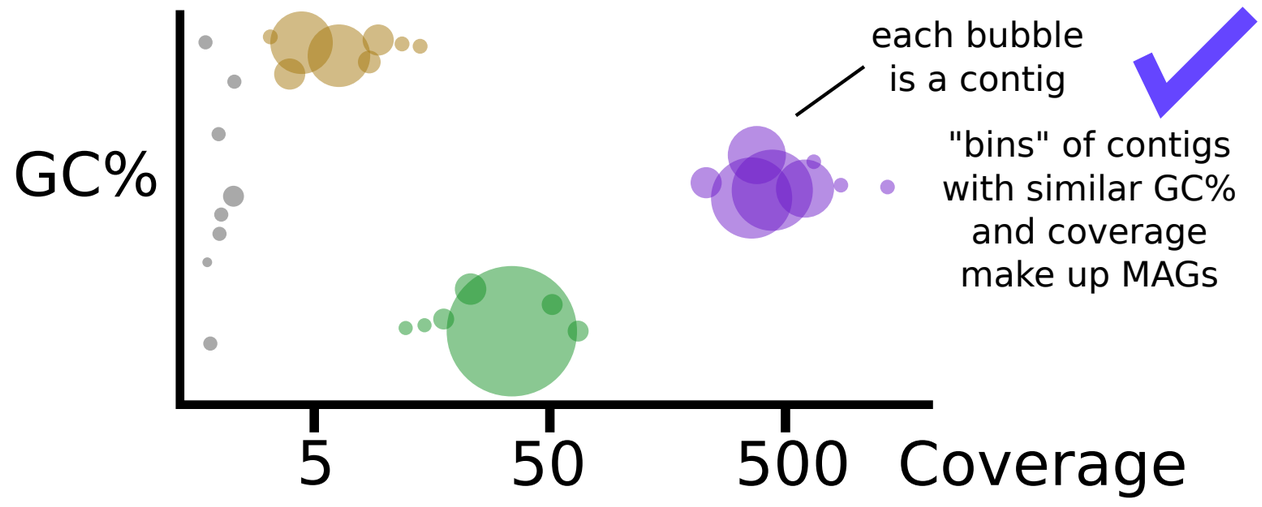
✗

For instance...
consider another hypothetical sample of equal abundance of two different species, but the genome of one is twice the size of the other



The chart is misleading because the difference between the two species is due to difference in the size of the genomes, not difference in abundance of the species.

Instead, the coverage can be determined from mapping the reads to the assembled contigs. This mostly captures what was seen with the amplicon sequencing.



✓

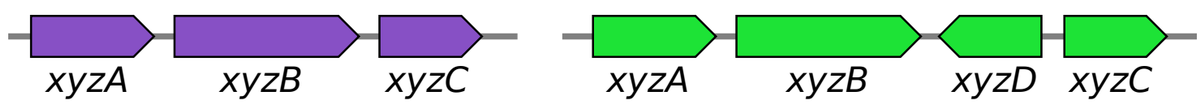
From this, bulk stats of the MAGs can be derived to give an overview of the genome size and the coverage.

	Coverage	Genome size	Total reads
spA	500x	3Mb	15M
spB	50	7Mb	3.5M
spC	5	4Mb	0.2M

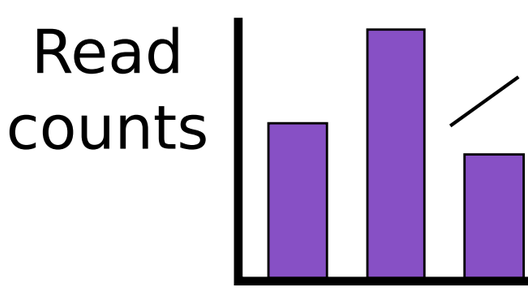
* this assumes a read length of 100bp

✓

Gene presence, absence, or individual operons could also be reconstructed and displayed.



In metagenomics, raw read counts within an operon will not be uniform, and will vary by length of the gene



✗

Normalized counts are uninformative, since they will just show the coverage. Transcriptomes are needed!

	spA	spB	spC
xyz	*ABC*	*ABDC*	—

Instead, consider a simple presence-absence table

✓