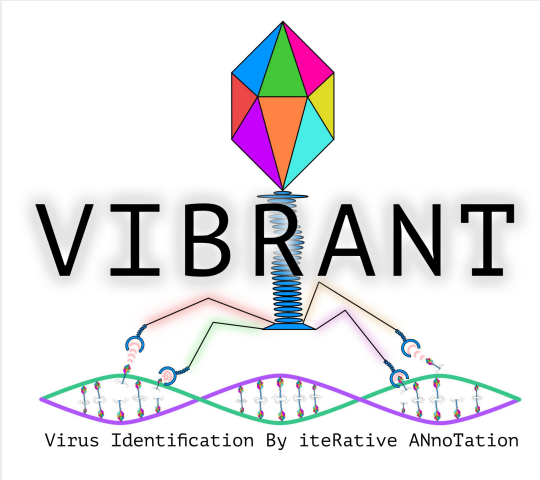
VIBRANT
file descriptions

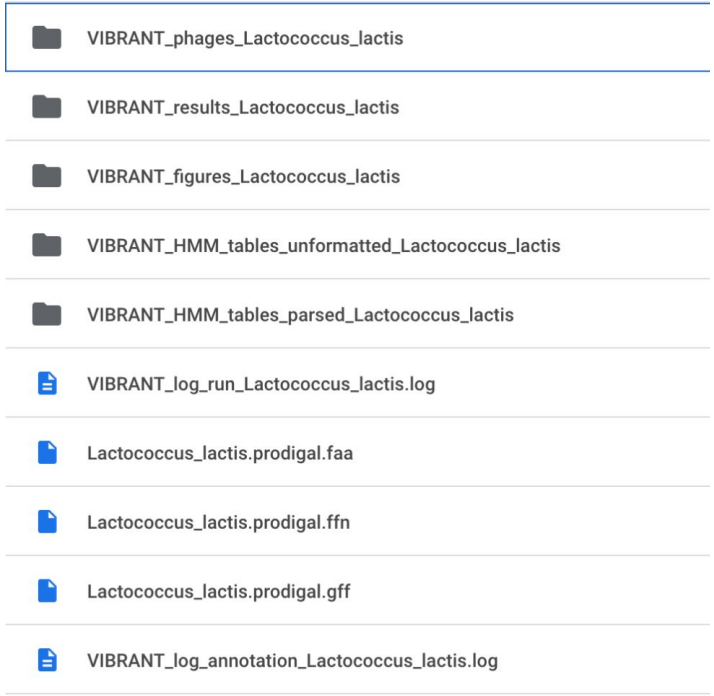


NORBIS Bioinformatics Summer School
Summer 2021

Anantharaman Lab, University of Wisconsin-Madison, USA

Tutorial

Overview of output folder



The main output folder contains 5 folders (next slides)

1. Phages
2. Results
3. Figures
4. HMM unformatted tables
5. HMM parsed tables

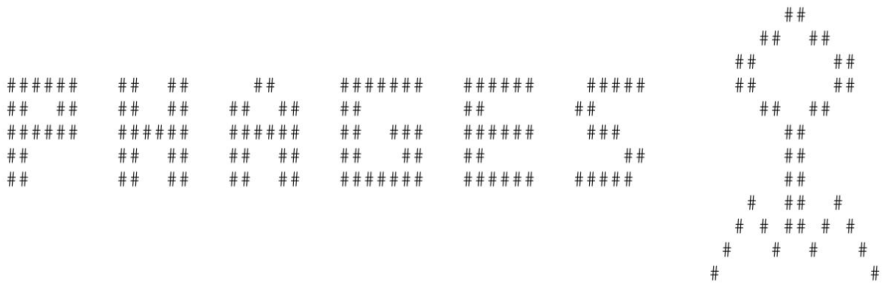
It also contains:

- **Log:** information about run time, how many sequences were analyzed
- **Annotation.log:**
- **Prodigal.faa** : nucleotides are translated to amino acids, the amino acids are saved here, in amino acid format
- **Prodigal.ffn** : nucleotides are translated to amino acids, the amino acids (their nucleotide CDS) are saved here
- **Prodigal.gff:** the GFF annotation files describing the translated amino acids.

The log files: VIBRANT_log_run_{samplename}.log

```
Command:  /home/mlangwig/miniconda3/envs/vibrant_env/bin/VIBRANT_run.py -i Lactococcus_lactis.fasta -t 2 -folder
Lactococcus_lactis.fasta_vibrant_folder
Date:     2021-08-10
Start:    12:37:29
End:      12:44:43
Runtime:  7.2 minutes
Program:  VIBRANT v1.2.1


1 scaffolds were read in.
1 scaffolds met minimum requirements: at least 1000bp and 4 ORFs.
5 putative phages were identified.
```

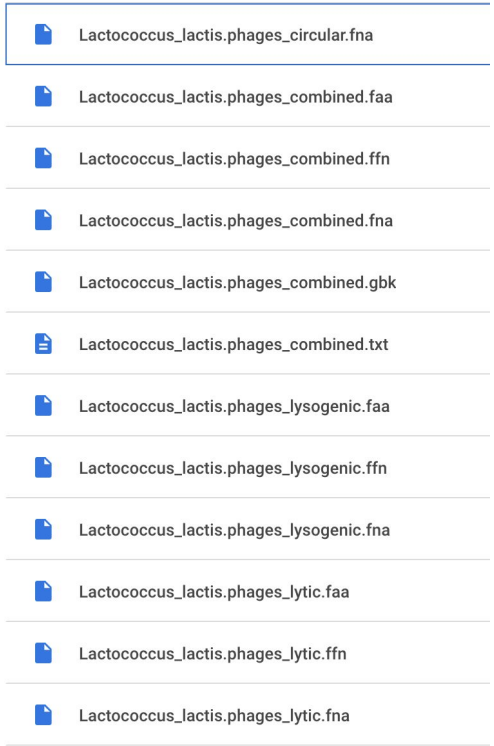


Note:
The default parameters of VIBRANT analyses scaffolds >1000bp and with 4 ORFs

The log files: VIBRANT_log_annotation_{samplename}.log

This file will be empty if there are no errors during annotation!

Folder 1: VIBRANT_phages_{filename}



This folder mostly contains text files related to the nucleotides or amino acids sequences.

VIBRANT identifies 4 types of phages:

- Lytic
- Combined
- Lysogenic
- Circular

For each of the phage types identified, there are 3 file formats: fna, faa and genbank.

(***.fna***, ***.faa***) are fasta format (starts with > header followed by either AGTC or amino acids)

There are also genbank files (***.gbk)*** which contains information in addition to a DNA/protein sequence

Folder 2: VIBRANT_results_{filename}

- VIBRANT_AMG_pathways_Lactococcus_lactis.tsv
- VIBRANT_figure_PCA_Lactococcus_lactis.tsv
- VIBRANT_integrated_prophage_coordinates_Lactococcus_lactis.tsv
- VIBRANT_machine_Lactococcus_lactis.tsv
- VIBRANT_complete_circular_Lactococcus_lactis.tsv
- VIBRANT_annotations_Lactococcus_lactis.tsv
- VIBRANT_AMG_counts_Lactococcus_lactis.tsv
- VIBRANT_summary_results_Lactococcus_lactis.tsv
- VIBRANT_genome_quality_Lactococcus_lactis.tsv
- VIBRANT_AMG_individuals_Lactococcus_lactis.tsv
- VIBRANT_summary_normalized_Lactococcus_lactis.tsv
- VIBRANT_genbank_table_Lactococcus_lactis.tsv

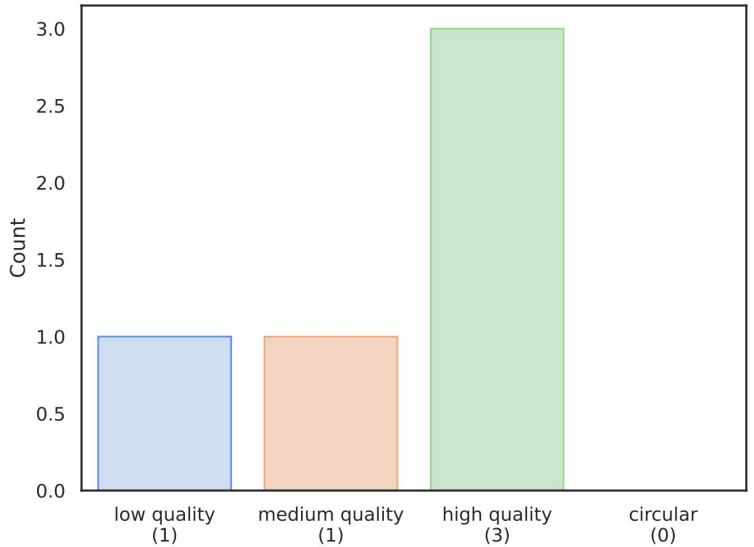
This folder contains TAB SEPARATED FILES (.tsv) which are text files separated by tabs. You can open them with R, Excel, or a simple text editor.

Most of these files contain detailed information about the scaffolds, what annotations were found, etc.

They are also used in generating figures (next slide).

Folder 3: VIBRANT_figures_{filename}

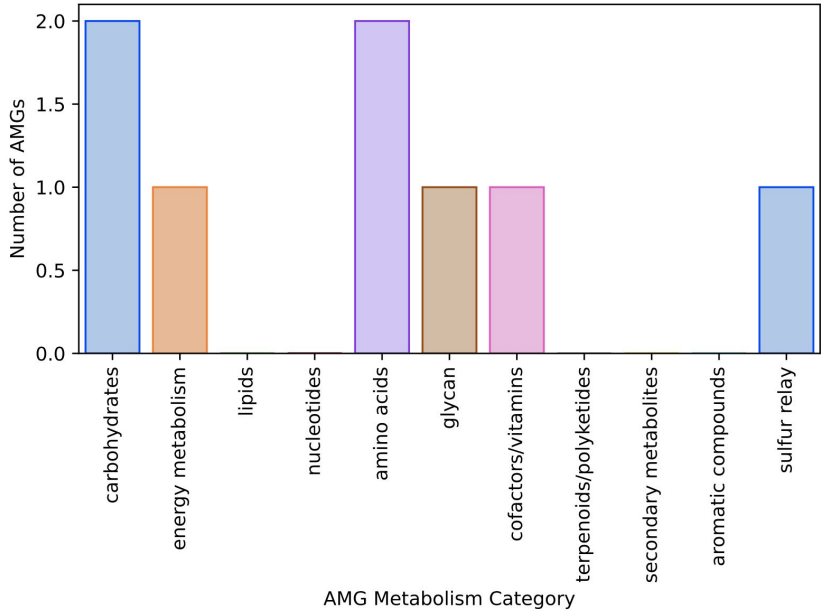
VIBRANT_figure_quality_Lactococcus_lactis.pdf	1:09 PM	11 KB
VIBRANT_figure_pathways_Lactococcus_lactis.pdf	1:09 PM	13 KB
VIBRANT_figure_sizes_Lactococcus_lactis.pdf	1:09 PM	10 KB
VIBRANT_figure_PCA_Lactococcus_lactis.pdf	1:09 PM	16 KB



How many of each quality type
Circular = most complete and highest quality

Folder 3: VIBRANT_figures_{filename}

VIBRANT_figure_quality_Lactococcus_lactis.pdf	1:09 PM	11 KB
VIBRANT_figure_pathways_Lactococcus_lactis.pdf	1:09 PM	13 KB
VIBRANT_figure_sizes_Lactococcus_lactis.pdf	1:09 PM	10 KB
VIBRANT_figure_PCA_Lactococcus_lactis.pdf	1:09 PM	16 KB



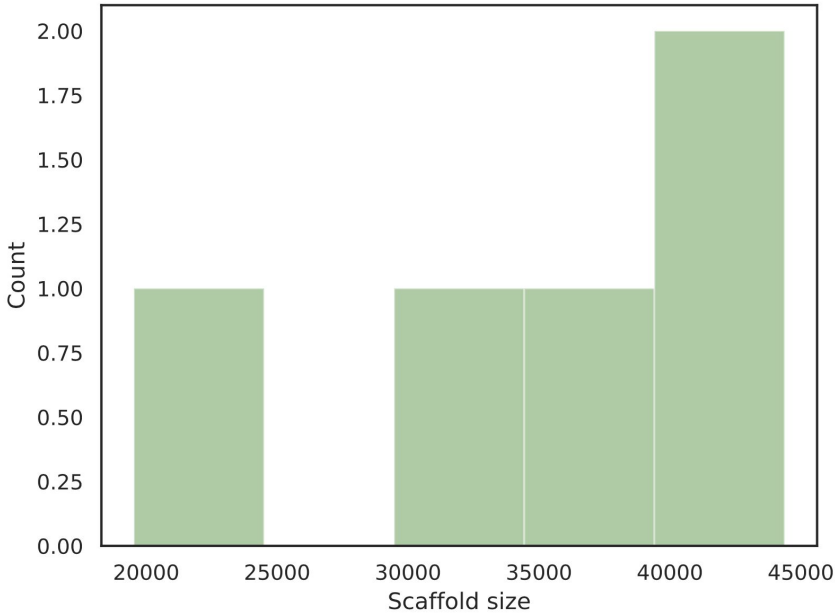
AMGs (auxiliary metabolic genes): genes acquired by viruses from hosts, that can tap into rate-limiting steps of host metabolism during infection.

They are useful to interpret in which processes these phages found contribute to.

Folder 3: VIBRANT_figures_{filename}

VIBRANT_figure_quality_Lactococcus_lactis.pdf	1:09 PM	11 KB
VIBRANT_figure_pathways_Lactococcus_lactis.pdf	1:09 PM	13 KB
VIBRANT_figure_sizes_Lactococcus_lactis.pdf	1:09 PM	10 KB
VIBRANT_figure_PCA_Lactococcus_lactis.pdf	1:09 PM	16 KB

Of the phages that are found, what are the scaffold sizes?
In this example, most scaffolds are between 40 000 and 45 000 bp long



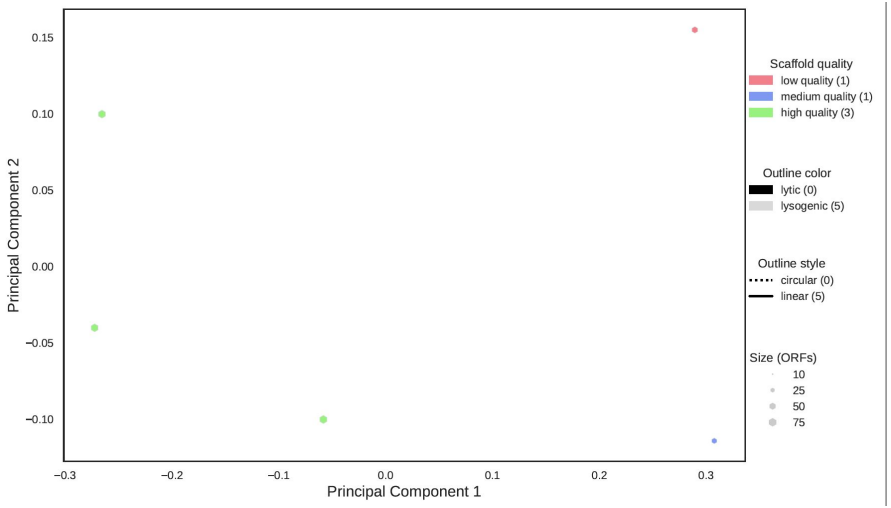
Folder 3: VIBRANT_figures_{filename}

VIBRANT_figure_quality_Lactococcus_lactis.pdf	1:09 PM	11 KB
VIBRANT_figure_pathways_Lactococcus_lactis.pdf	1:09 PM	13 KB
VIBRANT_figure_sizes_Lactococcus_lactis.pdf	1:09 PM	10 KB
VIBRANT_figure_PCA_Lactococcus_lactis.pdf	1:09 PM	16 KB

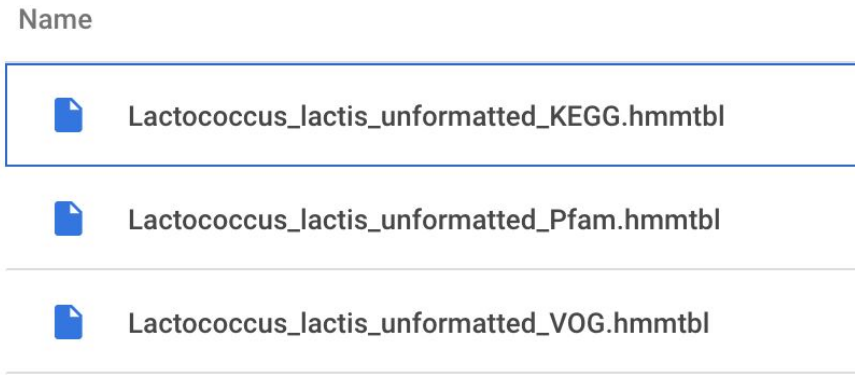
A **principal component analysis (PCA)** of the phages identified (in this example, there were 5 phages found)

Uses information about these scaffolds to compare them among each other.

Two points close to each other are more similar than if they were far apart. Axes are dimensionless. A PCA is a common statistical method that can be used for exploratory analyses. The points are labelled according to the legend.



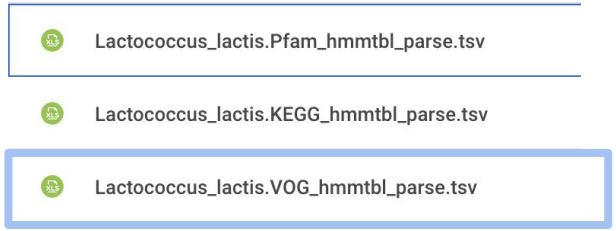
Folder 4: VIBRANT_HMM_tables_unformatted_{samplename}



When VIBRANT runs the HMM profiles on KEGG, Pfam, and VOG profiles, the results are written here.

You could use these files to parse out anything of interest to your study/research question.

Folder 5: VIBRANT_HMM_tables_parsed_{samplename}



Example:

protein	id	evalue	score
NC_002662.1$~	VOG22086	0	1582.8
NC_002662.1$~	VOG00844	0	1077.3
NC_002662.1$~	VOG00844	0	1032.8
NC_002662.1$~	VOG02857	0	1192.3
NC_002662.1$~	VOG04056	0	2338.1
NC_002662.1$~	VOG00179	7.10E-198	655
NC_002662.1$~	VOG13470	1.10E-172	566.7
NC_002662.1$~	VOG05830	1.30E-168	557
NC_002662.1$~	VOG01852	9.80E-155	511.8
NC_002662.1$~	VOG02979	1.70E-146	484.1
NC_002662.1$~	VOG17246	3.70E-134	441.5
NC_002662.1$~	VOG04273	1.70E-128	423.4

When VIBRANT runs the HMM profiles on KEGG, Pfam, and VOG profiles, the results are written here.

The column means:

- **Protein:** fasta header associated with the sequence
- **Id:** VOG, PFam, or KEGG identifier in the respective database
- **E-value:** Same meaning as in other bioinformatics settings. Standardized definition. Smallest value - best hit possible.
- **Score:** HMM score, analogous to a bit score in BLAST. The higher the number, the “better” the hit.

Your turn!

Discussion questions during the breakout room:

Answer these 3 questions about your scenario:

- How many sequences are circular, lytic, lysogenic, and combined phages?
- What is the distribution of genome quality?
- What are the 2 most prominent AMG categories in your example? Can you hypothesize why?

Use the rest of the time in the breakout room to discuss phages and VIBRANT questions with your assigned instructor. Do you have any questions to bring up as a group after the breakout room?