VIBRANT v1.2.0 Output Explanations

This document contains explanations for each of the files and folders that VIBRANT may generate. The first part contains a numbered list with short explanations of each file or folder. On the last page is a screenshot of an example output that references each number.

Key:

* Only present with nucleotide input
AMG: Auxiliary Metabolic Gene (i.e., virus-encoded metabolic gene)

Files/Folders

1. Parent output folder with "example" referring to the name of the input file.
2. * Prodigal predicted proteins for all input scaffolds.
3. * Prodigal predicted genes for all input scaffolds.
4. * Prodigal prediction gff file for all input scaffolds.
5. Folder containing PDF figures summarizing outputs:
6. Graphical summary (bar plot) of KEGG pathways for identified viral AMGs. File may not be present.
7. Graphical summary (nested bubble plot) for ratio of total input scaffolds (outside), number of scaffolds that were greater than or equal to minimum size restrictions (middle), and number of identified viruses (inside). Must have at least 10 input sequences for this file to be present. The represented numbers can be found in the log file (#22).
8. Graphical summary (bar plot) of number of viruses per genome quality category. Possible x-axis categories are high, medium, low and fragment quality draft, and *complete circular. Any category may not be present if no viruses were identified for that category.
9. * Graphical summary (histogram) of genome sizes of identified viruses.
10. Folder containing parsed HMM table raw outputs. Any one of the contained files may be empty. This folder is likely not of use but contains non-redundant annotation information. Contains both virus and non-virus annotations:
11. KEGG parsed HMM table raw outputs.
12. Pfam parsed HMM table raw outputs.
13. VOG parsed HMM table raw outputs.
14. Folder containing non-parsed (unformatted) HMM complete raw outputs. Any one of the contained files may be empty. This folder is likely not of use but contains complete annotation information. Contains both virus and non-virus annotations:
15. KEGG non-parsed HMM complete raw outputs.
16. Pfam non-parsed HMM complete raw outputs.
17. VOG non-parsed HMM complete raw outputs.
18. Log file. Contains information about the command used, the total runtime, date of run,

version of VIBRANT, and summary of values used in file #7. This log file is more of a run and output summary file.

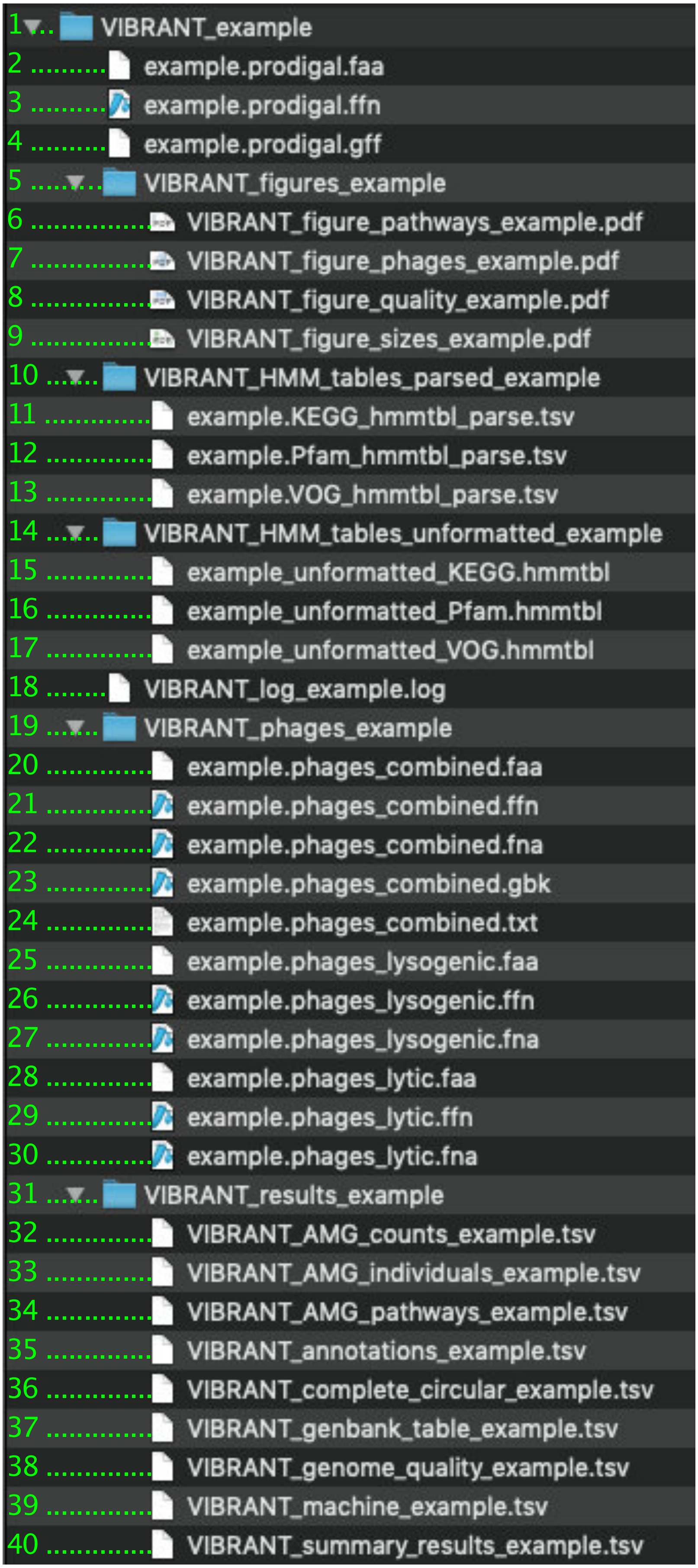
19. Folder containing FASTA and associated files for predicted viruses. Any one of the contained files may be empty if no viruses fit the criteria. Lysogenic viruses are determined by any virus scaffold excised from a larger scaffold or any that encodes an integrase. Lytic viruses are all others. For identified viruses:
20. All virus encoded proteins. *Note: any lysogenic virus that has been excised from a host scaffold will have the term "fragment" and a number appended to the original name to indicate that it does not represent the entire scaffold. This file, along with #25/#26/#27, will contain only the excised fragment.*
21. * All virus encoded genes.
22. * All virus genomes.
23. * GenBank format file for all virus genomes.
24. List of names (FASTA definition lines) for all virus genomes. *Note: This file may not entirely match the original input sequence names due to fragmentation of scaffolds. This list of names matches files #24/#25/#26/#27.*
25. Virus encoded proteins for predicted lysogenic viruses (subset of combined).
26. * Virus encoded genes for predicted lysogenic viruses (subset of combined).
27. * Virus genomes for predicted lysogenic viruses (subset of combined).
28. Virus encoded proteins for predicted lytic viruses (subset of combined).
29. * Virus encoded genes for predicted lytic viruses (subset of combined).
30. * Virus genomes for predicted lytic viruses (subset of combined).
31. Folder containing useful tab-delimited files for predicted viruses.
32. List of all predicted virus AMGs (by KEGG KO) and the total number of each. Summary of file #37. File may be empty.
33. List of individual predicted virus AMGs by protein and its respective genome. AMGs are determined by KEGG annotation but Pfam annotation is also given if applicable. See file #36 for summary. File may be empty.
34. List summarizing the present KEGG metabolic pathways (by KEGG map entry) corresponding to virus AMGs. See file #37 for individual AMGs. File may be empty. See this link for detailed information regarding KEGG metabolic pathways: https://www.genome.jp/kegg/pathway.html.
35. Complete list of annotations and associated information for KEGG, Pfam and VOG for all predicted viruses. Blank rows indicate proteins that were not given an annotation. Annotation names can be found in columns *KO/KO name*, *Pfam/Pfam name* and *VOG/VOG name*. Column *AMG* will indicate if the annotation was considered an "AMG" or not (blank). Columns for *evalue* and *score* refer to the annotation confidences generated by HMMsearch. Evalues are provided, but scores are used as the cutoff for annotations (must have a score of at least 40). Columns for *v-score* refer to the VIBRANT-specific "virus-like" score associated with each KO, Pfam and VOG. Briefly: scores of 0 indicate very low or no relatedness to viruses; 0.01 - 0.1 indicates low relatedness; 0.1 - 1 indicates moderate relatedness; 1 - 5 indicates significant relatedness; 5 - 10 indicates

substantial relatedness; 10 (max) for most cases indicates viral hallmark genes.

36. * Virus genomes that were predicted to be circular and therefore complete genomes. File may be empty. Circularization is determined by a kmer-based search between each end of the viral predicted genome. There must be at least a 20bp identical match.
37. List of the single annotation used for all predicted virus proteins. Annotation hierarchy for proteins annotated by multiple databases is as follows: KEGG > VOG > Pfam. This file is used to generate file #27 but will be present if #27 is not.
38. List summarizing the predicted genome quality and type (lytic/lysogenic) for all predicted viruses. Qualities may be fragment, low, medium or high quality draft. * Complete circular genomes, if applicable, are listed at the end and are redundant. That is, any complete circular genome will also be given a quality.
39. List summarizing predictions made by the neural network machine learning classifier. Will contain both viruses and non-viruses. This file is likely not of use but can be informative for checking outputs. There are curation steps following the classifier to validate predictions, so predictions in this file may not exactly match the final output. File may be empty if the classifier was not used.
40. List of complete annotation summary metrics for each predicted virus genome. The most useful metrics will be columns 1 through 8. All metrics shown, if applicable, were used for the neural network machine learning classifier. Explanations of each column:
    - scaffold: the name of the predicted virus
    - total genes: total number of predicted open reading frames
    - all KEGG: total number of KEGG annotations
    - KEGG v-score: sum of all KEGG annotation v-scores
    - all Pfam: total number of Pfam annotations
    - Pfam v-score: total number of Pfam annotations
    - all VOG: total number of VOG annotations
    - VOG v-score: total number of VOG annotations
    - KEGG int-rep: total number of KEGG integration related annotations (e.g., integrase, replicase, transposase). Useful for separating plasmids, mobile genetic elements and viruses
    - KEGG zero: total number of KEGG annotations that had a v-score of zero
    - Pfam int-rep: total number of Pfam integration related annotations (e.g., integrase, replicase, transposase). Useful for separating plasmids, mobile genetic elements and viruses
    - Pfam zero: total number of KEGG annotations that had a v-score of zero
    - VOG redoxin: total number of VOG redoxin related annotations (e.g., glutaredoxin, thioredoxin). Useful for separating viruses from bacteria/archaea because redoxins are common amongst viruses and therefore are given a high v-score, but are also common amongst bacteria/archaea
    - VOG rec-tran: total number of VOG integration related annotations that are not integrase (e.g., replicase, transposase). Useful for separating plasmids, mobile genetic elements and viruses

- VOG int: total number of VOG integrase related annotations. Used to identify putative lysogenic viruses. Also useful for separating plasmids, mobile genetic elements and viruses
- VOG RnR: total number of VOG ribonucleotide reductase (RnR) related annotations. Useful for separating viruses from bacteria/archaea because RnRs are common amongst viruses and therefore are given a high v-score, but are also common amongst bacteria/archaea
- VOG DNA: total number of VOG nucleotide (DNA/RNA) replication related annotations. Useful for separating viruses from bacteria/archaea because nucleotide replication proteins are common amongst viruses and therefore are given a high v-score, but are also common amongst bacteria/archaea. This is also a metric used for predicting viral completeness (along with VOG special) because genome replication is an essential process
- KEGG restriction: total number of KEGG restriction nuclease related annotations. Useful for separating plasmids, mobile genetic elements and viruses
- KEGG toxin: total number of KEGG toxin/anti-toxin related annotations. Useful for separating plasmids, mobile genetic elements and viruses
- VOG special: total number of VOG hallmark annotations (e.g., virion structural proteins, holin/lysin, terminase). This is a metric used for predicting viral completeness (along with VOG DNA)
- annotation check: the number of proteins annotated by KEGG, Pfam and VOG
- p_v check: the number of proteins annotated by Pfam and VOG only
- p_k check: the number of proteins annotated by KEGG and Pfam only
- k_v check: the number of proteins annotated by KEGG and VOG only
- k check: the number of proteins annotated by KEGG only
- p check: the number of proteins annotated by Pfam only
- v check: the number of proteins annotated by VOG only
- h check: the number of proteins not annotated

See next page for respective screenshot of file/folder outputs:



```
1 ▼.. VIBRANT_example
2 .......... example.prodigal.faa
3 .......... example.prodigal.ffn
4 .......... example.prodigal.gff
5 ....▼... VIBRANT_figures_example
6 ............... VIBRANT_figure_pathways_example.pdf
7 ............... VIBRANT_figure_phages_example.pdf
8 ............... VIBRANT_figure_quality_example.pdf
9 ............... VIBRANT_figure_sizes_example.pdf
10 ...▼.. VIBRANT_HMM_tables_parsed_example
11 ............. example.KEGG_hmmtbl_parse.tsv
12 ............. example.Pfam_hmmtbl_parse.tsv
13 ............. example.VOG_hmmtbl_parse.tsv
14 ...▼.. VIBRANT_HMM_tables_unformatted_example
15 ............. example_unformatted_KEGG.hmmtbl
16 ............. example_unformatted_Pfam.hmmtbl
17 ............. example_unformatted_VOG.hmmtbl
18 ........ VIBRANT_log_example.log
19 ...▼.. VIBRANT_phages_example
20 ............. example.phages_combined.faa
21 ............. example.phages_combined.ffn
22 ............. example.phages_combined.fna
23 ............. example.phages_combined.gbk
24 ............. example.phages_combined.txt
25 ............. example.phages_lysogenic.faa
26 ............. example.phages_lysogenic.ffn
27 ............. example.phages_lysogenic.fna
28 ............. example.phages_lytic.faa
29 ............. example.phages_lytic.ffn
30 ............. example.phages_lytic.fna
31 ...▼.. VIBRANT_results_example
32 ............. VIBRANT_AMG_counts_example.tsv
33 ............. VIBRANT_AMG_individuals_example.tsv
34 ............. VIBRANT_AMG_pathways_example.tsv
35 ............. VIBRANT_annotations_example.tsv
36 ............. VIBRANT_complete_circular_example.tsv
37 ............. VIBRANT_genbank_table_example.tsv
38 ............. VIBRANT_genome_quality_example.tsv
39 ............. VIBRANT_machine_example.tsv
40 ............. VIBRANT_summary_results_example.tsv
```