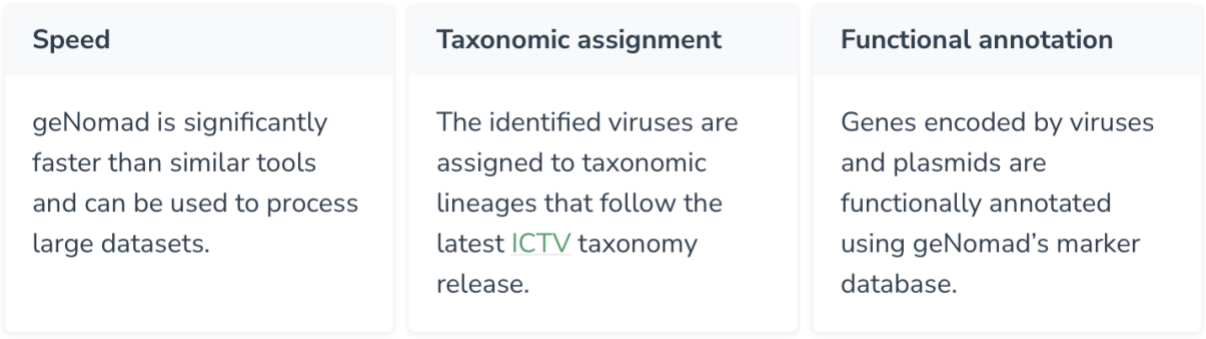
Viruses & Plasmids Workflow (v1.0)



Overview

This workflow takes in assembly files, generates a list of viruses and plasmids that were detected in the file, and provides quality and confidence information.

Running the Workflow

Currently, this workflow can be run in NMDC EDGE or on local compute resources. (Installation instructions and requirements are found here and here.)

Tutorial videos on how to run each workflow in NMDC EDGE are found here.

Input

The input for this workflow must be an assembly file from a metagenome, metatranscriptome, or genome assembly workflow. The recommended input is the output from the NMDC metagenome assembly or metatranscriptome workflow.

- **Acceptable file formats:** .fasta, .fa, .fna

Details

This workflow takes in assembly files and runs the geNomad workflow, followed by checkV to determine the quality and confidence of the geNomad results. The taxonomy that is reported is based on the ICTV guidelines. A quickstart guide for geNomad can be found here.

Software Versions

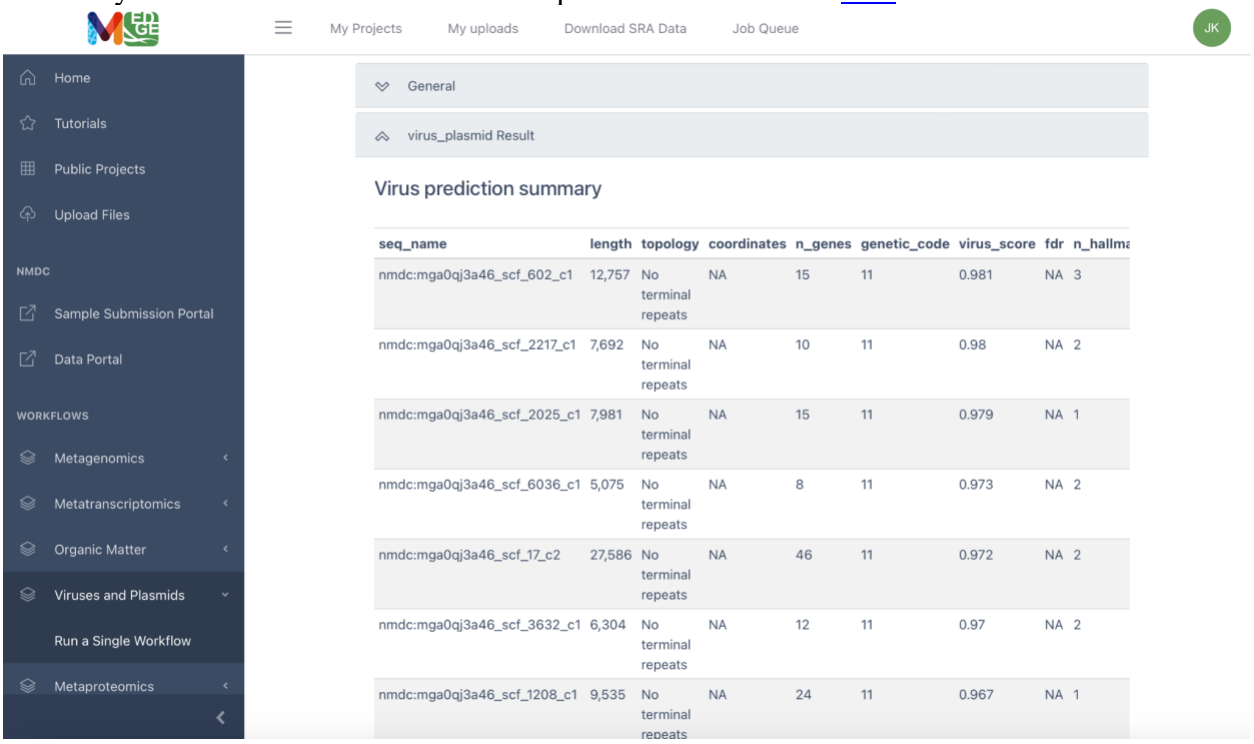
- geNomad: v.1.5.2
- geNomad database: v1.3
- CheckV: v1.0.1
- CheckV database: v1.4

Default parameters: minimum score: 0.7, at least one hallmark gene identified for short contigs.

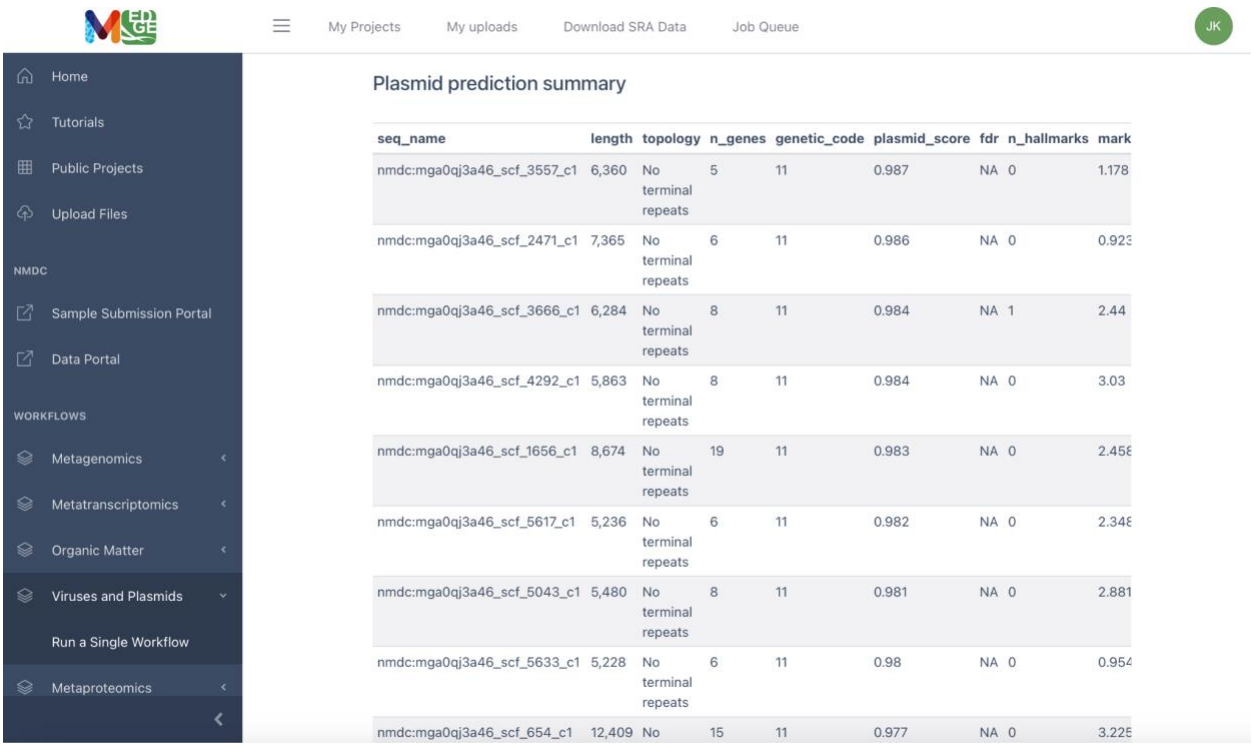
Relaxed parameters: will report all sequences that are identified as "virus" or "plasmid", regardless of the score itself or any other annotation; “relaxed” setting minimum score: 0, no requirements for the identification of hallmark genes.
Conservative parameters: minimum score: 0.8, identification of at least one hallmark gene is required for all contigs.

Output

In NMDC EDGE, the virus_plasmid result tab displays information about predicted viruses in the input data including sequence length, topology, coordinates, number of genes, genetic code, virus score, false discovery rate (FDR), number of hallmark genes, marker enrichment, and taxonomy. More information about this output data can be found here.

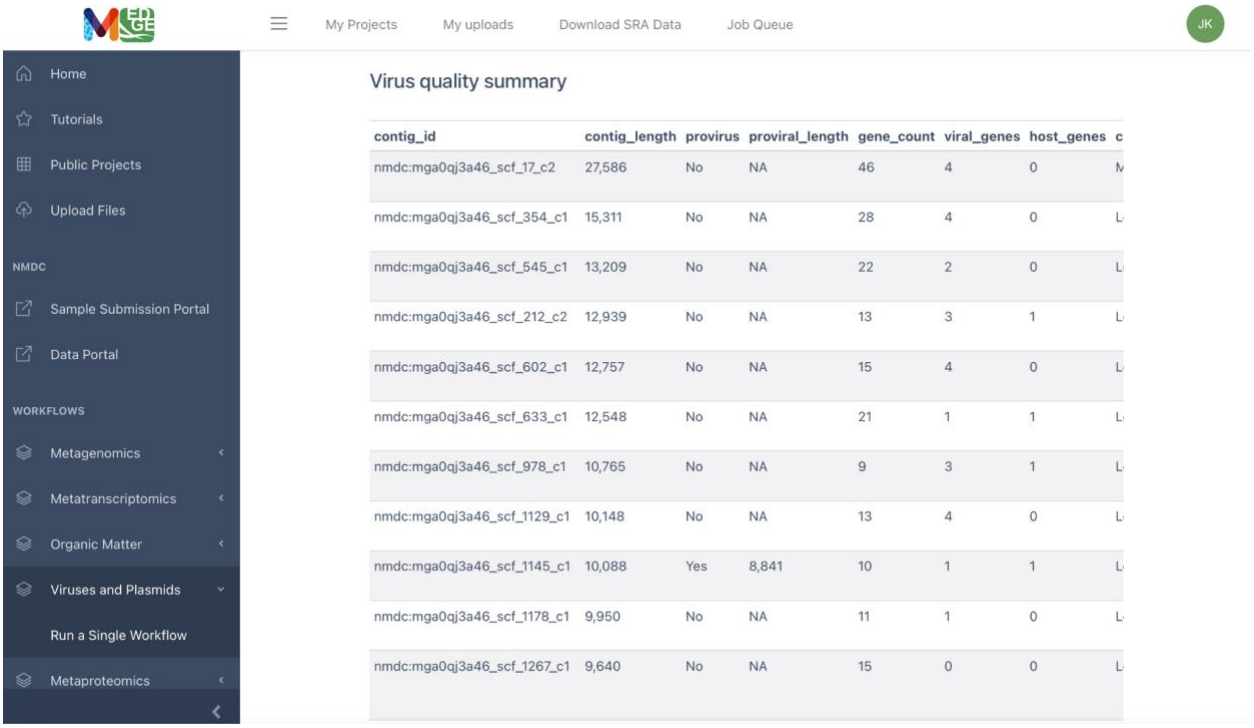


Another table in this section provides the plasmid prediction summary which includes information on sequence length, topology, number of genes, genetic code, plasmid score, false discovery rate (FDR), number of hallmark genes, marker enrichment, conjugation genes, and any antimicrobial resistance (AMR) genes present. As stated above, more information on this output data can be found here.



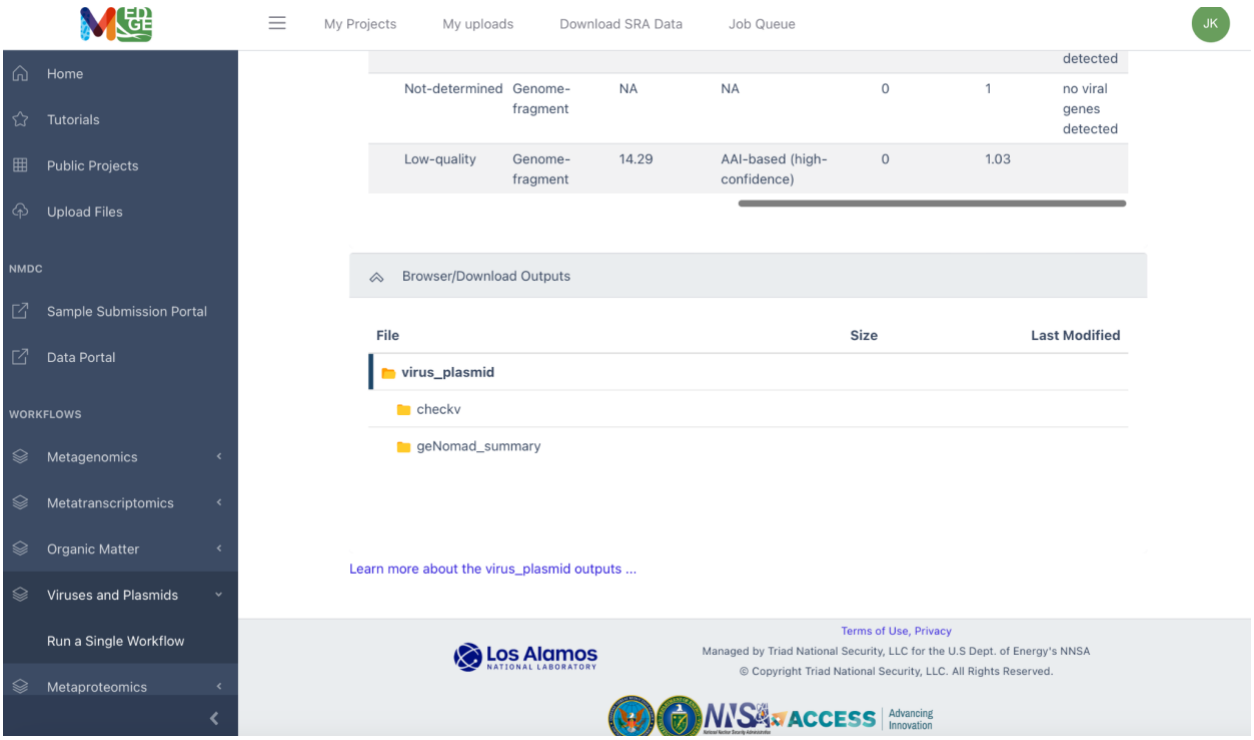
seq_name	length	topology	n_genes	genetic_code	plasmid_score	fdr	n_hallmarks	mark
nmdc:mga0qj3a46_scf_3557_c1	6,360	No terminal repeats	5	11	0.987	NA	0	1.178
nmdc:mga0qj3a46_scf_2471_c1	7,365	No terminal repeats	6	11	0.986	NA	0	0.923
nmdc:mga0qj3a46_scf_3666_c1	6,284	No terminal repeats	8	11	0.984	NA	1	2.44
nmdc:mga0qj3a46_scf_4292_c1	5,863	No terminal repeats	8	11	0.984	NA	0	3.03
nmdc:mga0qj3a46_scf_1656_c1	8,674	No terminal repeats	19	11	0.983	NA	0	2.458
nmdc:mga0qj3a46_scf_5617_c1	5,236	No terminal repeats	6	11	0.982	NA	0	2.348
nmdc:mga0qj3a46_scf_5043_c1	5,480	No terminal repeats	8	11	0.981	NA	0	2.881
nmdc:mga0qj3a46_scf_5633_c1	5,228	No terminal repeats	6	11	0.98	NA	0	0.954
nmdc:mga0qj3a46_scf_654_c1	12,409	No	15	11	0.977	NA	0	3.225

A virus quality summary table is also provided, where it details the contig ID, contig length, provirus information, gene counts, quality information, completeness information, completeness method, contamination, kmer frequency, and any relevant warnings.



contig_id	contig_length	provirus	proviral_length	gene_count	viral_genes	host_genes	c
nmdc:mga0qj3a46_scf_17_c2	27,586	No	NA	46	4	0	M
nmdc:mga0qj3a46_scf_354_c1	15,311	No	NA	28	4	0	L
nmdc:mga0qj3a46_scf_545_c1	13,209	No	NA	22	2	0	L
nmdc:mga0qj3a46_scf_212_c2	12,939	No	NA	13	3	1	L
nmdc:mga0qj3a46_scf_602_c1	12,757	No	NA	15	4	0	L
nmdc:mga0qj3a46_scf_633_c1	12,548	No	NA	21	1	1	L
nmdc:mga0qj3a46_scf_978_c1	10,765	No	NA	9	3	1	L
nmdc:mga0qj3a46_scf_1129_c1	10,148	No	NA	13	4	0	L
nmdc:mga0qj3a46_scf_1145_c1	10,088	Yes	8,841	10	1	1	L
nmdc:mga0qj3a46_scf_1178_c1	9,950	No	NA	11	1	0	L
nmdc:mga0qj3a46_scf_1267_c1	9,640	No	NA	15	0	0	L

All output files are available to download under the Browser/Download Outputs tab at the bottom of the results page.


My Projects
My uploads
Download SRA Data
Job Queue
JK
Home
Tutorials
Public Projects
Upload Files
NMDC
Sample Submission Portal
Data Portal
WORKFLOWS
Metagenomics
Metatranscriptomics
Organic Matter
Viruses and Plasmids
Run a Single Workflow
Metaproteomics
detected
Not-determined Genome-fragment NA NA 0 1 no viral genes detected
Low-quality Genome-fragment 14.29 AAI-based (high-confidence) 0 1.03
Browser/Download Outputs
File
Size
Last Modified
virus_plasmid
checkv
geNomad_summary
Learn more about the virus_plasmid outputs ...
Terms of Use, Privacy
Managed by Triad National Security, LLC for the U.S Dept. of Energy's NNSA
© Copyright Triad National Security, LLC. All Rights Reserved.
Los Alamos NATIONAL LABORATORY
ACCESS
Advancing Innovation