Workflow de Virus y Plásmidos (v1.0)

Speed	Taxonomic assignment	Functional annotation
geNomad is significantly faster than similar tools and can be used to process large datasets.	The identified viruses are assigned to taxonomic lineages that follow the latest ICTV taxonomy release.	Genes encoded by viruses and plasmids are functionally annotated using geNomad's marker database.

Descripción General

Este workflow toma archivos de ensamblaje, genera una lista de virus y plásmidos que se detectaron en el archivo y proporciona información de calidad y confianza.

Ejecutando el Workflow

Actualmente, este workflow se puede ejecutar en NMDC EDGE o en recursos informáticos locales. (Las instrucciones y requisitos de instalación se encuentran aquí y aquí).

Aquí se encuentran videos tutoriales sobre cómo ejecutar cada workflow en NMDC EDGE.

Aporte

El input para este workflow debe ser un archivo de 'reads' ensamblados provenientes de un metagenoma, metatranscriptoma o genoma. La entrada recomendada es la salida del 'Metagenomic Assembly Workflow' del NMDC (Ensamblaje Metagenómica) o del workflow 'metaT' (Metatranscriptómica).

- **Formatos de archivo aceptables:** .fasta, .fa, .fna

Detalles

Este workflow toma archivos de ensamblaje y ejecuta el workflow geNomad, seguido de checkV para determinar la calidad y confianza de los resultados de geNomad. La taxonomía que se reporta se basa en los linajes del ICTV. Puede encontrar una guía de inicio rápido para geNomad aquí.

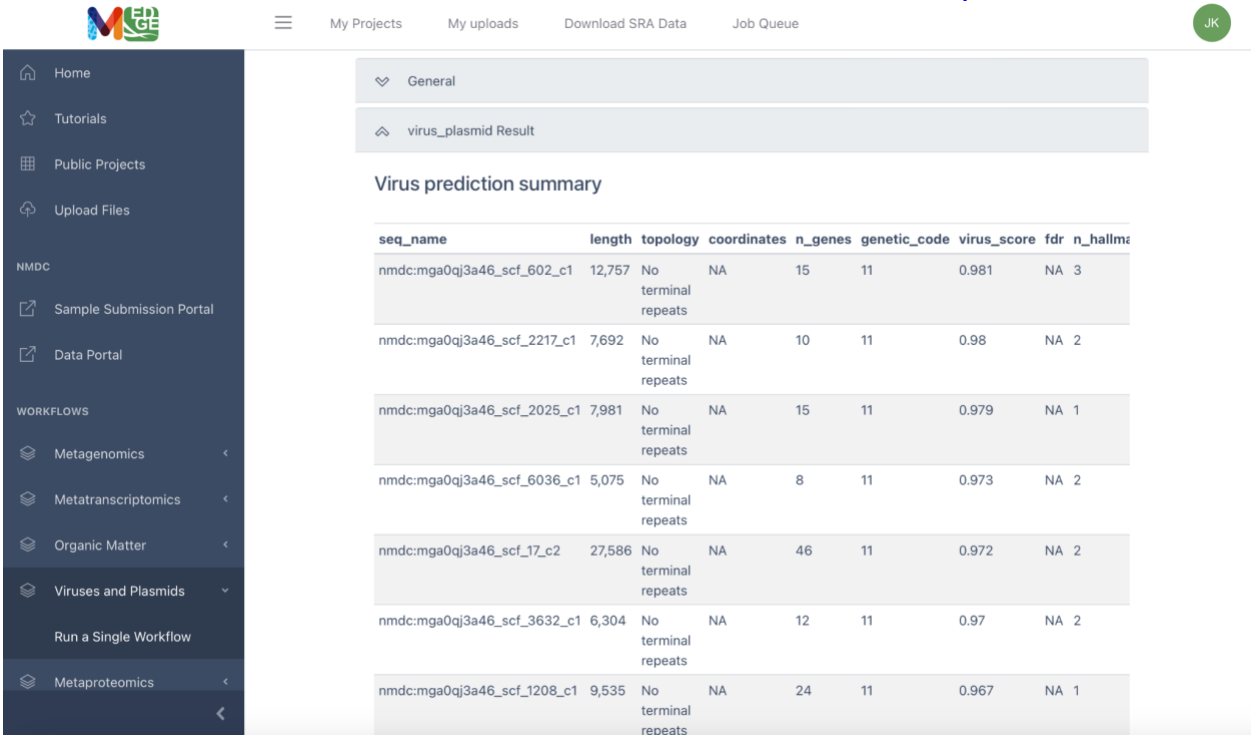
Versiones de Software

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

Parámetros predeterminados: puntuación mínima: 0,7, al menos un gen distintivo identificado para ‘contigs’ cortos.
Parámetros relajados: informará todas las secuencias identificadas como ‘virus’ o ‘plásmido’, independientemente de la puntuación en sí o de cualquier otra anotación; Puntuación mínima de configuración ‘relajada’: 0, sin requisitos para la identificación de genes distintivos
Parámetros conservadores: puntuación mínima: 0,8; se requiere la identificación de al menos un gen viral distintivo para todos los ‘contigs’

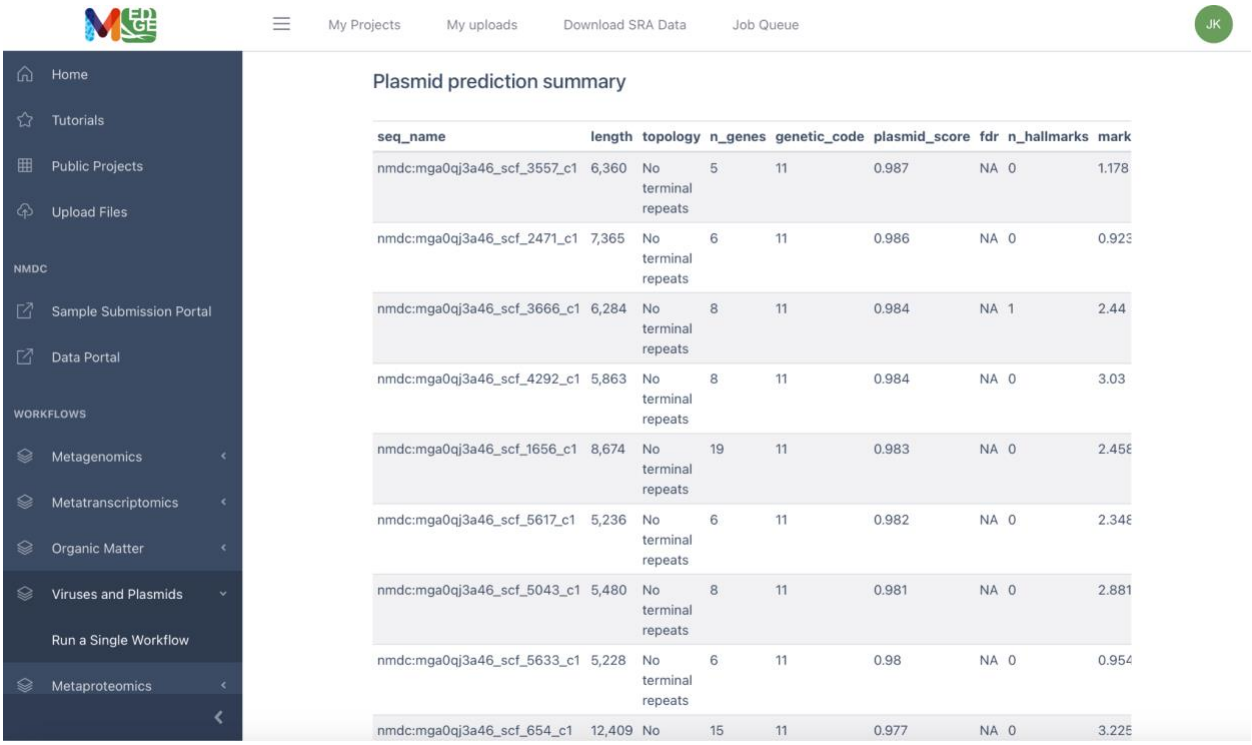
Productos

En NMDC EDGE, la tabulación de resultados virus_plasmid muestra información sobre los virus identificados en los datos de entrada, incluyendo la longitud de la secuencia, la topología, las coordenadas, la cantidad de genes, el código genético, la puntuación del virus, la tasa de descubrimiento falso (FDR), la cantidad de genes distintivos, el enriquecimiento del marcador, y taxonomía. Puede encontrar más información sobre estos datos de salida aquí.



seq_name	length	topology	coordinates	n_genes	genetic_code	virus_score	fdr	n_hallma
nmdc:mga0qj3a46_scf_602_c1	12,757	No terminal repeats	NA	15	11	0.981	NA	3
nmdc:mga0qj3a46_scf_2217_c1	7,692	No terminal repeats	NA	10	11	0.98	NA	2
nmdc:mga0qj3a46_scf_2025_c1	7,981	No terminal repeats	NA	15	11	0.979	NA	1
nmdc:mga0qj3a46_scf_6036_c1	5,075	No terminal repeats	NA	8	11	0.973	NA	2
nmdc:mga0qj3a46_scf_17_c2	27,586	No terminal repeats	NA	46	11	0.972	NA	2
nmdc:mga0qj3a46_scf_3632_c1	6,304	No terminal repeats	NA	12	11	0.97	NA	2
nmdc:mga0qj3a46_scf_1208_c1	9,535	No terminal repeats	NA	24	11	0.967	NA	1

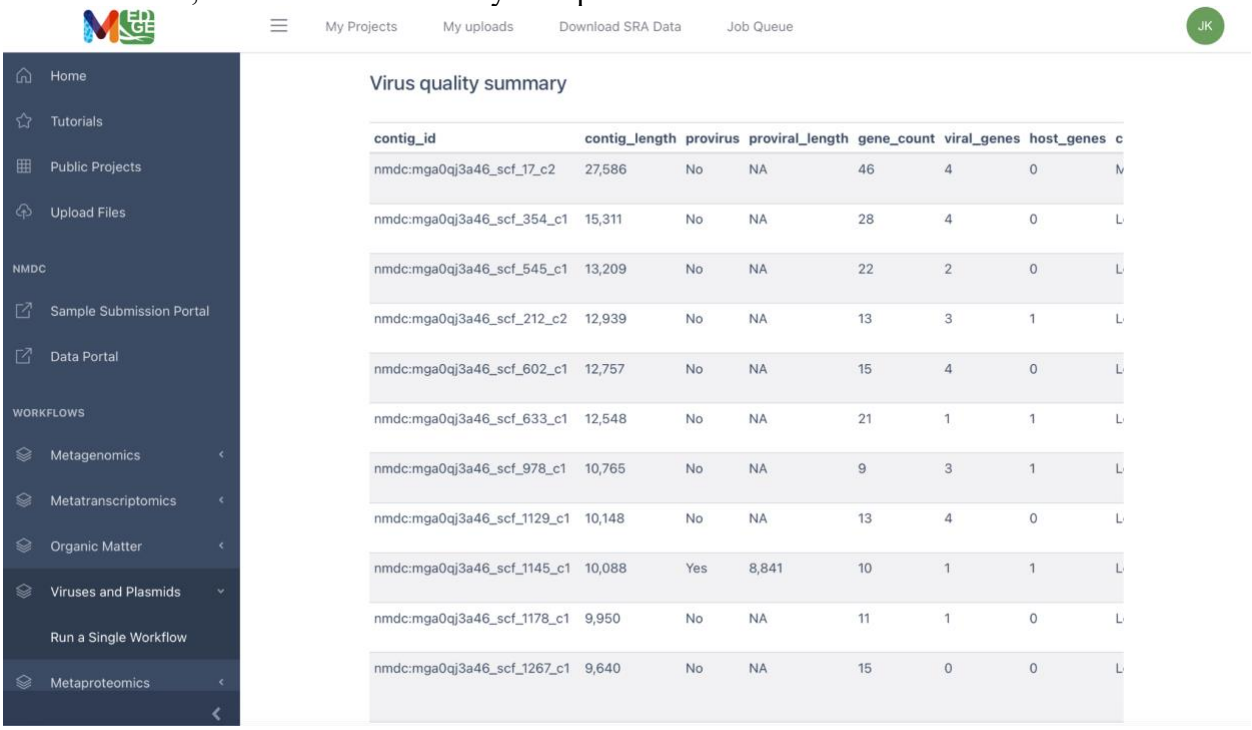
Otra tabla en esta sección proporciona el resumen de predicción de plásmidos que incluye información sobre la longitud de la secuencia, la topología, la cantidad de genes, el código genético, la puntuación del plásmido, la tasa de descubrimiento falso (FDR), la cantidad de genes distintivos, el enriquecimiento de marcadores, los genes de conjugación y cualesquiera genes de resistencia antimicrobiana (AMR) presentes. Como se indicó anteriormente, puede encontrar más información sobre estos datos de salida aquí.



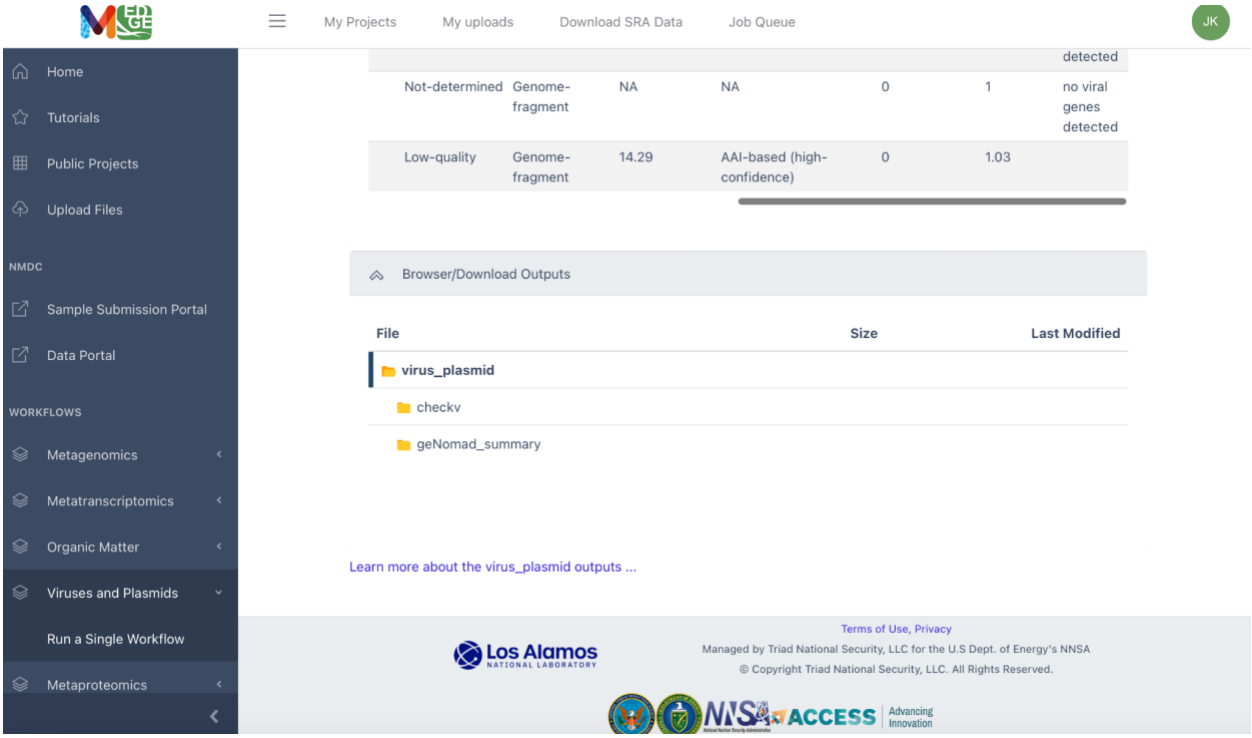
Plasmid prediction summary

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

También se proporciona una tabla de resumen de la calidad del virus, donde se detalla la identificación del 'contig', la longitud del 'contig', la información del provirus, el recuento de genes, la información de calidad, la información de integridad, el método de integridad, la contaminación, la frecuencia de kmer y cualquier advertencia relevante.



Virus quality summary

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

Todos los archivos de salida están disponibles para descargar en la tabulación ‘Browser/Download Output’ (Navegador/Descargar Productos) en la parte inferior de la página de resultados.