

## Workflow « geNomad » pour la détection de virus et plasmides (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 <a href="#">ICTV</a> taxonomy release.	Genes encoded by viruses and plasmids are functionally annotated using geNomad's marker database.

### Aperçu

Ce workflow prend en entrée les contigs obtenus après assemblage, génère une liste de virus et plasmides détectés dans l'assemblage, et fournit des informations sur la qualité et la confiance de ces prédictions.

### Exécuter ce workflow

Ce workflow peut être exécuté via [NMDC EDGE](#) ou sur des ressources de calcul locales (les instructions et conditions d'installation se trouvent [ici](#) et [ici](#).)

Des didacticiels vidéo sur la façon d'exécuter chaque workflow dans NMDC EDGE sont disponibles [ici](#).

### Fichiers d'entrée

L'entrée de ce workflow doit être un fichier de reads assemblés (i.e. contigs) provenant d'un workflow d'assemblage de métagénome, de métatranscriptome ou de génome. L'entrée recommandée est la sortie du workflow d'assemblage de métagénome ou de métatranscriptome NMDC.

- Formats de fichier acceptés : .fasta, .fa, .fna

### Instructions détaillées

Ce workflow accepte en entrée les fichiers de séquences assemblées et exécute l'outil geNomad, suivi de l'outil checkV pour déterminer la qualité et la fiabilité des résultats de geNomad. La taxonomie rapportée par geNomad est basée sur les dernières [recommandations de l'ICTV](#). Un guide de démarrage rapide pour geNomad est disponible [ici](#).

### Versions des outils

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

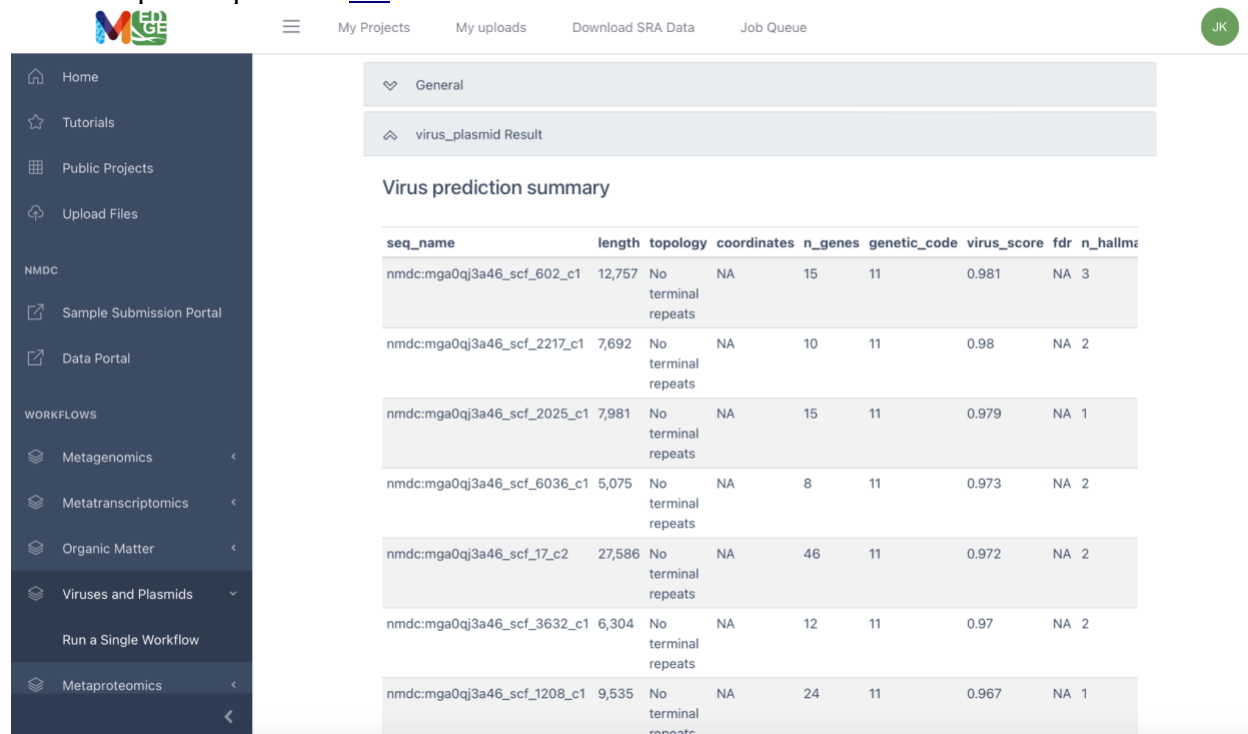
**Paramètres par défaut:** score minimum: 0.7, au moins un gène marqueur de virus identifié pour les contigs courts.

**Paramètres relâchés (“relaxed”):** le fichier de résultat inclura toutes les séquences identifiées comme « virus » ou « plasmide », quel que soit le score lui-même ou toute autre annotation ; Score minimum du paramètre relâché: 0, aucun nombre minimum de gène marqueur.

**Paramètres stricts (“conservative”):** score minimum: 0.8, au moins un gène marqueur de virus identifié pour tous les contigs.

## Fichiers de sortie

Dans NMDC EDGE, l'onglet de résultats « virus\_plasmid » affiche des informations sur les virus prédits dans les données d'entrée, notamment la longueur de la séquence, la topologie (linéaire ou circulaire), les coordonnées (si détection d'un provirus), le nombre de gènes, le code génétique, le score du virus, un taux estimé de fausses découvertes (FDR), le nombre de gènes marqueurs, l'enrichissement en gènes marqueurs, et la taxonomie. Plus d'explications sur ces résultats spmt disponibles [ici](#).



seq_name	length	topology	coordinates	n_genes	genetic_code	virus_score	fdr	n_hallmarks
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

Un autre tableau de cette section fournit le résumé des prédictions de plasmides, et comprend des informations sur la longueur de la séquence, la topologie, le nombre de gènes, le code génétique, le score « plasmide », un taux estimé de fausses découvertes (FDR), le nombre de gènes caractéristiques, l'enrichissement en marqueurs, les gènes de conjugaison et gènes de résistance (AMR) présents. Comme indiqué ci-dessus, plus d'informations sur ces données de sortie peuvent être trouvées [ici](#).

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

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

Un tableau récapitulatif de la qualité des prédictions de virus est également fourni, qui comprend l'identifiant du contig, la longueur du contig, les informations sur le (pro)virus, le nombre de gènes, les informations sur la qualité (estimation de la longueur totale du génome et comparaison avec la longueur de la séquence observée), les méthodes utilisées pour estimer la qualité, la contamination, la fréquence de kmer et tout autre potentiel problème de qualité identifié par CheckV.

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

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

Tous les fichiers de sortie peuvent être téléchargé sous l'onglet « Browser/Download Outputs » au bas de la page de résultats.

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

My Projects

My uploads

Download SRA Data

Job Queue

JK

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 ...

Los Alamos

NATIONAL LABORATORY

Managed by Triad National Security, LLC for the U.S Dept. of Energy's NNSA

© Copyright Triad National Security, LLC. All Rights Reserved.

MISA ACCESS

Advancing Innovation