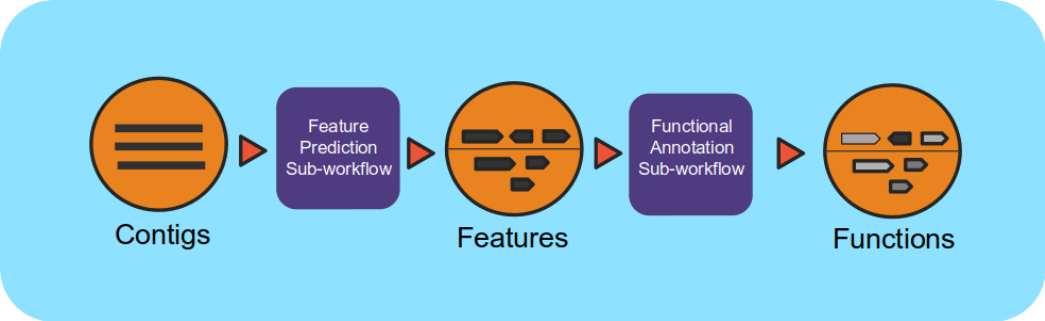
Workflow d'Annotation des Métagénomes (v1.0.0)



Aperçu

Ce workflow prend en entrée des métagénomes assemblés et génère des annotations structurelles et fonctionnelles.

Exécution du 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)

Des didacticiels vidéo sur la façon d'exécuter chaque worfklow dans NMDC EDGE sont disponibles ici.

Fichiers d'entrée

Ce workflow nécessite des contigs assemblés dans un fichier FASTA. Ce fichier d'entrée peut être le fichier de sortie d'un workflow d'assemblage de métagénomes.

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

Instructions détaillées

Le workflow utilise un certain nombre d'outils et de bases de données libres (open source) pour générer les annotations structurelles et fonctionnelles. L'assemblage fourni en entrée est d'abord divisé en fractions de 10 Mo pour être traitées en parallèle. Selon la configuration du moteur de workflow, le fractionnement peut être traité en parallèle. Chaque division est d'abord annotée structurellement, puis ces résultats sont utilisés pour l'annotation fonctionnelle. L'annotation structurelle utilise tRNAscan_se, RFAM, CRT, Prodigal et GeneMarkS. Ces résultats sont fusionnés pour créer une annotation structurelle par consensus. Le résultat de l'annotation structurelle (au format GFF) est utilisé pour l'annotation fonctionnelle qui utilise plusieurs bases de données de familles de protéines (SMART, COG, TIGRFAM, SUPERFAMILY, Pfam et Cath-FunFam) ainsi que des modèles HMM personnalisés. Les prédictions fonctionnelles sont créées à l'aide de Last et HMM. Ces annotations sont également fusionnées dans un fichier GFF de consensus. Enfin, les annotations fractionnées respectives sont fusionnées pour générer un seul fichier d'annotation structurelle et un seul fichier d'annotation fonctionnelle. De plus, plusieurs fichiers récapitulatifs sont générés au format TSV.

Versions des outils

- Conda
- tRNAscan-SE >= 2.0
- Infernal 1.1.2
- CRT-CLI 1.8
- Prodigal 2.6.3
- GeneMarkS-2 >= 1.07
- Last >= 983
- HMMER 3.1b2

LA-UR-21-21661

- TMHMM 2.0

Fichiers de sortie

Les principaux résultats sont le fichier d'annotation structurelle et le fichier d'annotation fonctionnelle. Le fichier d'annotation fonctionnelle peut être une entrée pour le workflow « MAGs Generation ».

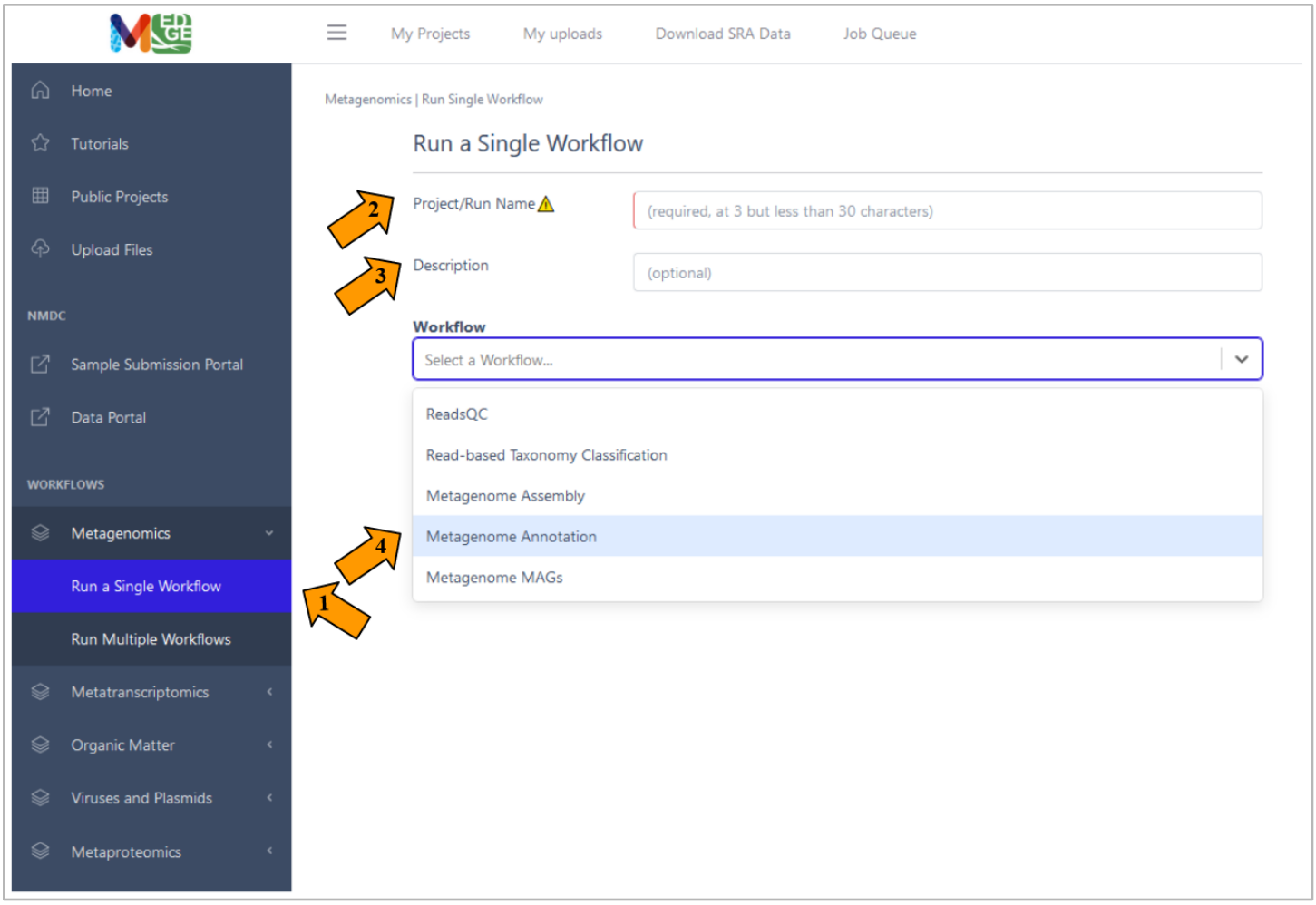
Fichiers de sortie principaux	Description
Structural Annotation	Fichier d'annotation structurelle basée sur plusieurs outils (.gff)
Functional Annotation	Fichier d'annotation fonctionnelle basée sur plusieurs outils (.gff)
KEGG summary	Résumé des affiliation fonctionnelles basées sur la base de données KEGG (.tsv)
EC summary	Résumé des affiliation fonctionnelles basées sur la base de données EC (Enzyme Commission) (.tsv)
Gene phylogeny summary	Résumé des affiliation taxonomiques des gènes (.tsv)

Exécution du workflow d'Annotation des Métagénomes « Metagenome Annotation » dans NMDC EDGE

Sélectionner un workflow

1. Dans la catégorie Metagenomics dans la barre de menu de gauche, sélectionnez 'Run a Single Workflow'.
2. Entrez un nom de projet ***unique*** sans espaces (les traits de soulignement sont possibles).
3. Une description est facultative, mais utile.
4. Sélectionnez « Metagenome Annotation » dans le menu déroulant sous Workflow.

LA-UR-21-21661

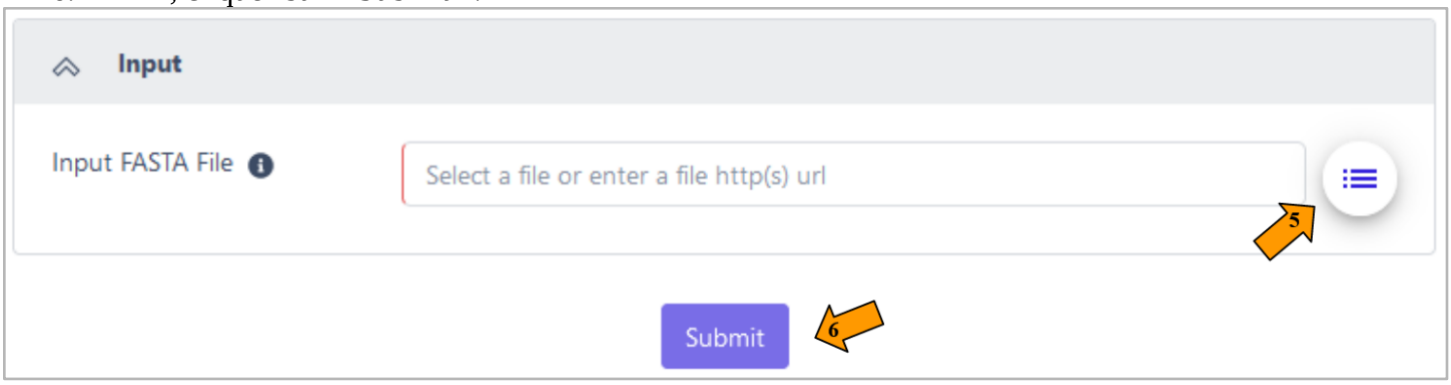


Fichiers d'entrée

Ce workflow accepte des données assemblées au format FASTA en entree; the file can be compressed. Il est recommande d'utiliser en fichier d'entrée les fichiers de sortie du workflow « Metagenome Assembly ».

Formats de fichiers acceptables: .fasta, .fa, .fasta.gz, .fa.gz, fna.gz.

5. Cliquez sur le bouton à droite de l'espace vide pour sélectionner le fichier de données a analyser. Une boîte de dialogue appelée « Select a File » s'ouvrira pour permettre à l'utilisateur de trouver le fichier souhaité à partir d'un workflow précédemment exécuté, du dossier de données publiques, ou d'un fichier téléchargé par l'utilisateur
6. Enfin, cliquez sur « Submit ».



LA-UR-21-21661

Fichiers de sortie

La section « General » indique quel workflow et quels outils ont été exécutés, ainsi que les informations d'exécution.

General

Workflow	Run	Status	Running Time	Start	End
Metagenome Annotation	On	Done	01:22:05	2021-10-14 15:07:49	2021-10-14 16:29:54

"Project Configuration" : {...}

La section « Metagenome Annotation Result » contient des statistiques sur les séquences traitées, les gènes prédits et les informations générales sur la qualité des résultats du workflow.

Metagenome Annotation Result

Processed Sequences Statistics

Data type	Number of seqs	Number of bps	Median length	Average length	Length shortest seq	Length longest seq	Standard deviation
final_fasta	25,726	52,201,077	818.5	2,029.118	200	859,644	16,939.403
sequences_with_genes	24,248	51,497,305	865	2,123.775	200	859,644	17,443.493
sequences_without_genes	1,478	703,772	404	476.165	203	1,918	217.554

Predicted Genes Statistics

Feature type	Prediction method	Number of seqs	Number of bps	Median length	Average length	Length shortest seq	Length longest seq	Standard deviation	Number of predicted features
CDS	Prodigal v2.6.3	12,478	3,694,932	180	228.831	75	1,935	156.372	16,147
CDS	GeneMark.hmm-2 v1.05	18,576	35,352,681	480	669.267	90	16,545	616.622	52,823
tRNA	tRNAscan-SE v.2.0.7 (Oct 2020)	451	67,404	76	79.486	56	146	10.062	848
misc_feature	INFERNAL 1.1.3 (Nov 2019)	4	1,454	366.5	363.5	349	372	10.408	4
regulatory	INFERNAL 1.1.3 (Nov 2019)	4	1,454	366.5	363.5	349	372	10.408	4
ncRNA	INFERNAL 1.1.3 (Nov 2019)	4	1,454	366.5	363.5	349	372	10.408	4
rRNA	INFERNAL 1.1.3 (Nov 2019)	4	1,454	366.5	363.5	349	372	10.408	4
tmRNA	INFERNAL 1.1.3 (Nov 2019)	4	1,454	366.5	363.5	349	372	10.408	4
CRISPR	CRT 1.8.2	11	7,170	456	551.538	155	1,168	341.877	13

General Quality Info

Name	Status
Coding density	74.88%
Genes per 1M bp	1,353.8
Seqs per 1M bp	492.83

La section « Browser/Download » fournit des fichiers de sortie disponibles au téléchargement. Les fichiers de sortie principaux sont l'annotation fonctionnelle et l'annotation structurelle (gff). Le fichier d'annotation fonctionnelle est requis pour le workflow de génération de MAGs avec les contigs assemblés.

Browser/Download Outputs

File	Size	Last Modified
MetagenomeAnnotation		
Annotation_Test.faa	20.53 MB	20 days ago
Annotation_Test_cath_funfam.gff	11.89 MB	20 days ago
Annotation_Test_cog.gff	7.92 MB	20 days ago
Annotation_Test_contigs.fna	51.30 MB	20 days ago
Annotation_Test_crt.crisprs	11 kB	20 days ago
Annotation_Test_ec.tsv	1.27 MB	20 days ago
Annotation_Test_functional_annotation.gff	17.43 MB	20 days ago
Annotation_Test_gene_phylogeny.tsv	10.45 MB	20 days ago
Annotation_Test_ko.tsv	2.36 MB	20 days ago
Annotation_Test_ko_ec.gff	44.29 MB	20 days ago
Annotation_Test_pfam.gff	9.71 MB	20 days ago
Annotation_Test_product_names.tsv	5.21 MB	20 days ago
Annotation_Test_proteins.cath_funfam.domtblout	151.86 MB	20 days ago
Annotation_Test_proteins.cog.domtblout	51.46 MB	20 days ago
Annotation_Test_proteins.pfam.domtblout	15.08 MB	20 days ago
Annotation_Test_proteins.smart.domtblout	7.59 MB	20 days ago
Annotation_Test_proteins.supfam.domtblout	339.68 MB	20 days ago
Annotation_Test_proteins.tigrfam.domtblout	3.00 MB	20 days ago
Annotation_Test_smart.gff	3.33 MB	20 days ago
Annotation_Test_structural_annotation.gff	9.99 MB	20 days ago
Annotation_Test_structural_annotation_stats.json	6 kB	20 days ago
Annotation_Test_structural_annotation_stats.tsv	3 kB	20 days ago
Annotation_Test_supfam.gff	12.60 MB	20 days ago
Annotation_Test_tigrfam.gff	1.79 MB	20 days ago
rc	2 B	20 days ago
script	35 kB	20 days ago

LA-UR-21-21661