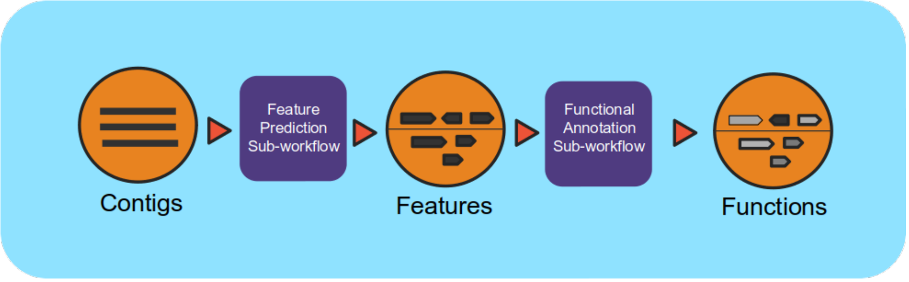
Metagenome Annotation Workflow (v1.0.0)


Contigs
Feature Prediction Sub-workflow
Features
Functional Annotation Sub-workflow
Functions

Overview

This workflow takes assembled metagenomes and generates structural and functional annotations.

Running the Workflow

Currently, this workflow can be run in NMDC EDGE or from the command line. (CLI instructions and requirements are found here.)

Input

Metagenome Annotation requires assembled contigs in a FASTA file. This input can be the output from the Metagenome Assembly workflow and this is recommended.

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

Details

The workflow uses a number of open-source tools and databases to generate the structural and functional annotations. The input assembly is first split into 10MB splits to be processed in parallel. Depending on the workflow engine configuration, the split can be processed in parallel. Each split is first structurally annotated, then those results are used for the functional annotation. The structural annotation uses tRNAscan_se, RFAM, CRT, Prodigal and GeneMarkS. These results are merged to create a consensus structural annotation. The resulting GFF is the input for functional annotation which uses multiple protein family databases (SMART, COG, TIGRFAM, SUPERFAMILY, Pfam and Cath-FunFam) along with custom HMM models. The functional predictions are created using Last and HMM. These annotations are also merged into a consensus GFF file. Finally, the respective split annotations are merged together to generate a single structural annotation file and single functional annotation file. In addition, several summary files are generated in TSV format.

Software Versions

- 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
- TMHMM 2.0

Output

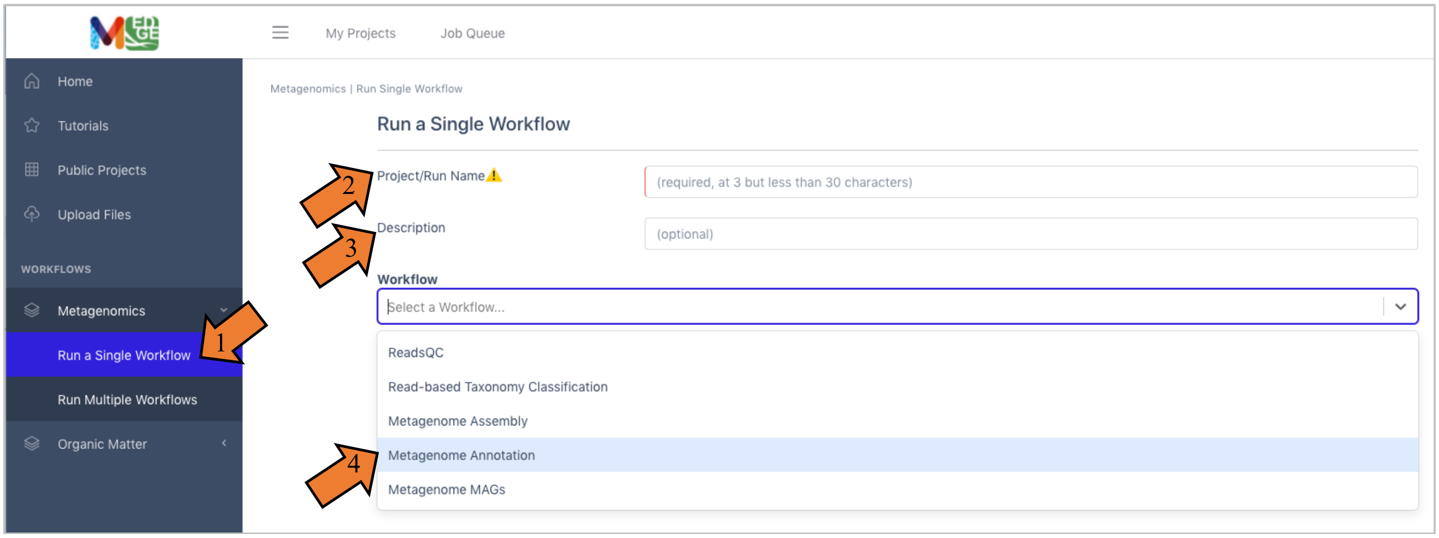
The main outputs are the structural annotation file and the functional annotation file. The functional annotation file can be an input for the MAGs Generation workflow.

Primary Output Files	Description
Structural Annotation	Consensus structural annotation file from multiple tools (.gff)
Functional Annotation	Consensus functional annotation file from multiple tools (.gff)
KEGG summary	KEGG gene function tabular summary (.tsv)
EC summary	Enzyme Commission tabular summary (.tsv)
Gene phylogeny summary	Gene phylogeny tabular summary (.tsv)

Running the Metagenome Annotation Workflow in NMDC EDGE

Select a workflow

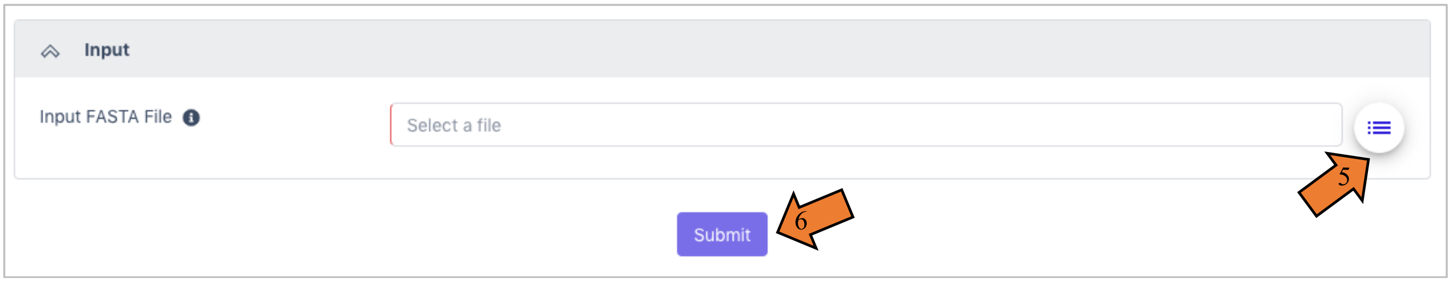
1. From the Metagenomics category in the left menu bar, select 'Run a Single Workflow'.
2. Enter a ***unique*** project name with no spaces (underscores are fine).
3. A description is optional, but helpful.
4. Select 'Metagenome Annotation' from the dropdown menu under Workflow.



Input

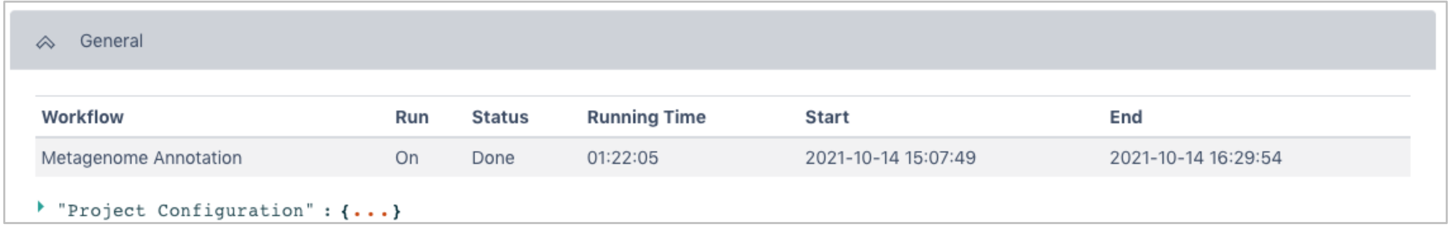
This workflow accepts assembled Illumina data in FASTA format as the input; the file can be compressed. (It is highly recommended to input the assembled contigs from the Metagenome Assembly workflow.) **Acceptable file formats:** .fasta, .fa, .fna, .fasta.gz, .fa.gz, .fna.gz.

5. Click the button to the right of the input blank for data to select the data file for the analysis. (If there are separate files, there will be two input blanks.) A box called 'Select a File' will open to allow the user to find the desired file(s) from previously run projects, the public data folder, or files uploaded by the user.
6. Then click 'Submit'.



Output

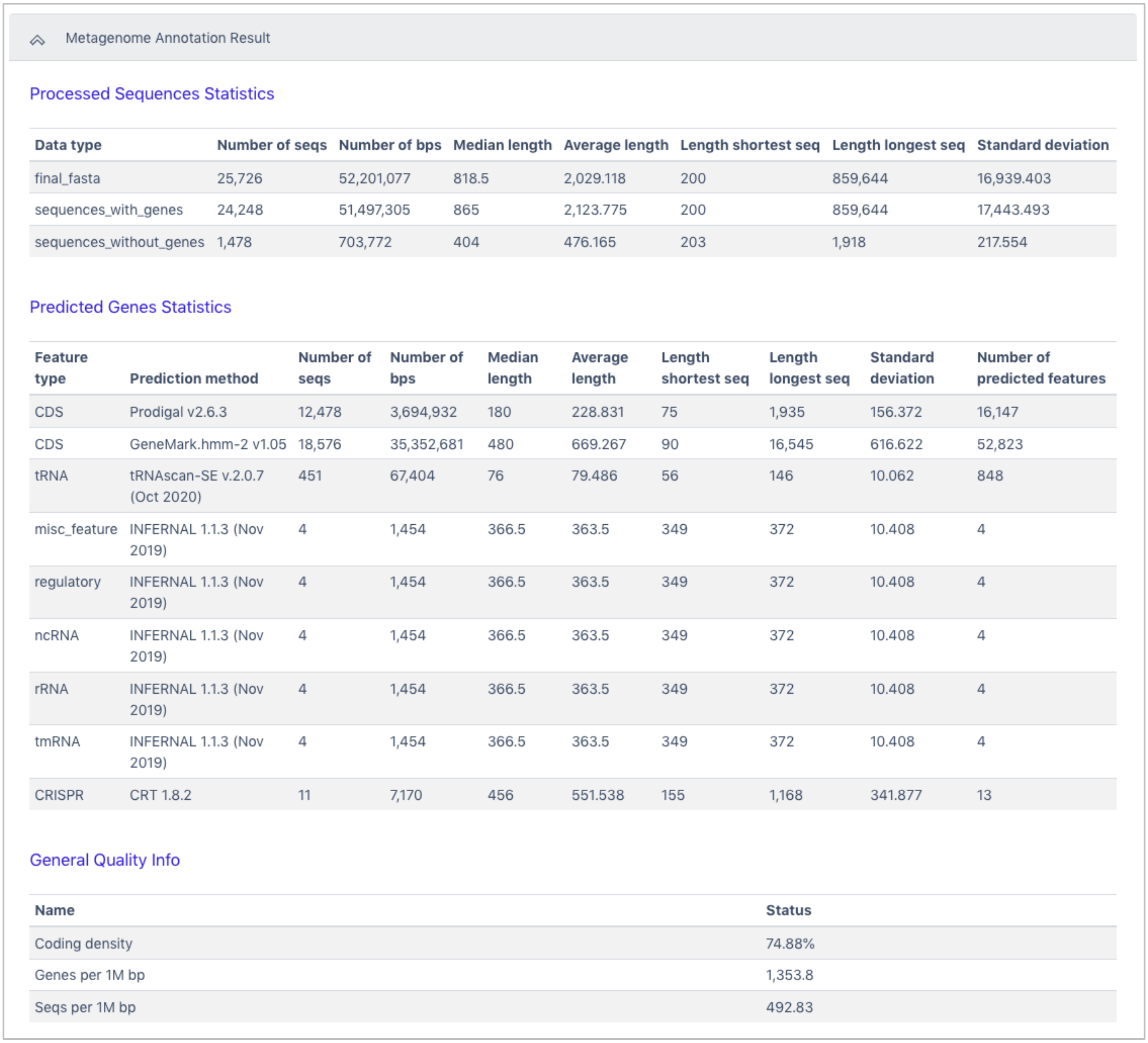
The General section of the output shows which workflow was run and the run time information.



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" : {...}
```

The Metagenome Annotation Result section has statistics for Processed Sequences, Predicted Genes, and General Quality Information from the 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

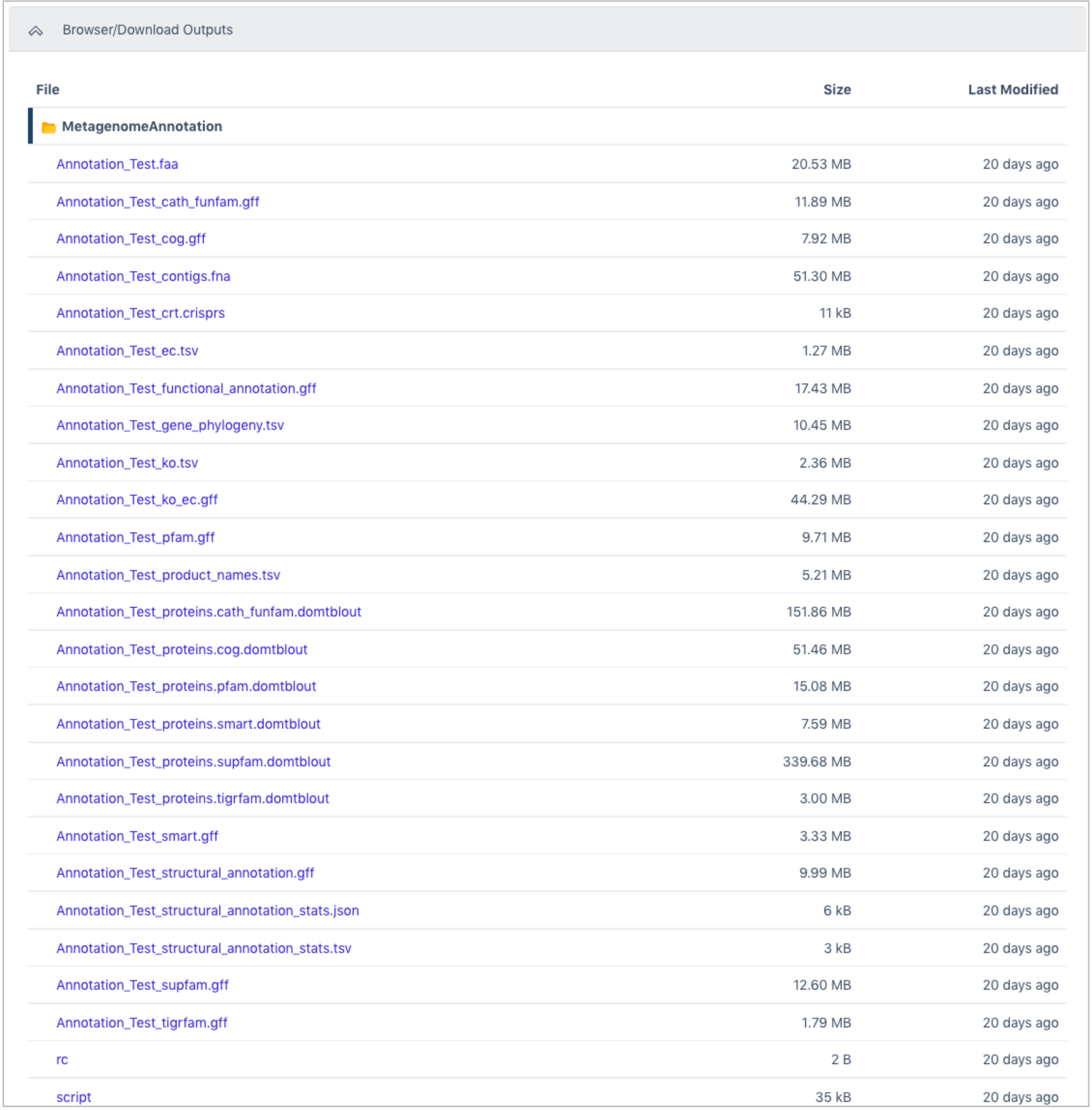
Predicted Genes Statistics

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

The Browser/Download Output section provides output files available to download. The primary results are the functional annotation and the structural annotation files (.gff). The functional annotation file is required input for the MAGs Generation workflow along with the assembled contigs.



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