MiDSystem Tutorial

Table of Contents

General Instruction

MiDSystem consists of four pipelines: reference-guided genome assembly, non-reference-guided assembly, long-read/hybrid assembly, and metagenomics analysis.

Reference-guided genome assembly and non-reference-guided assembly pipelines are for the single microbial species with short-read sequencing data (e.g., Illumina). If you already have a reference genome for your assembled sequences, we suggest conducting the reference-guided pipeline. In addition, MiDSystem provides assembly pipelines of the single microbial species with long-read sequences (e.g., Pacific Biosciences and Oxford Nanopore Technologies). Also, the hybrid assembly allows performing genome assembly with short-read and long-read sequences simultaneously. Before submission, please make sure what kind of experimental design in the uploaded files and select a suitable pipeline for analysis. Trying to run metagenomics samples on these pipelines above, or to run single species data on the metagenomics pipeline will fail the analysis process since they are designed for different purposes. The pipelines menu locates at the top of the home page (Fig. 1).

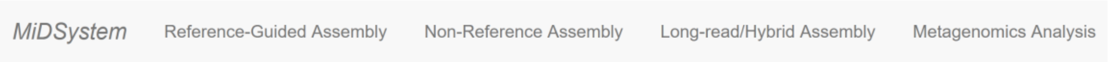
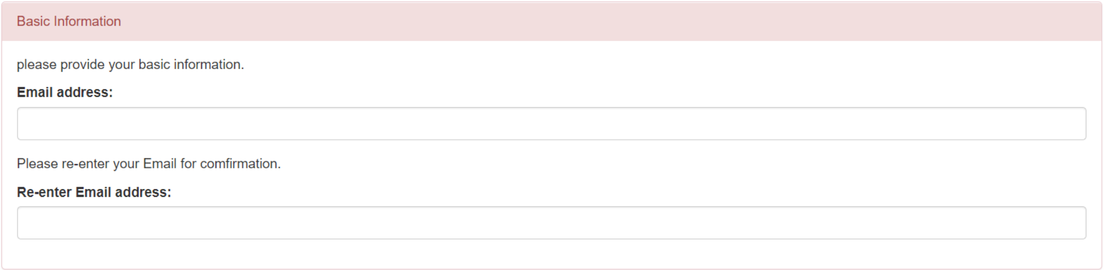


Fig. 1. The menu at the top of the home page allows users to select the pipelines to run.

There are step-by-step instructions on each pipeline's home page. Please follow them. For all the pipelines, you have to input your basic information first (i.e., the email address) (Fig. 2), so that MiDSystem can send a unique link to check the status or the final report of the submitted task. The unique link for your task is generated for security reasons. With the correct email address and link, you can come back to view your data at any time, so feel free to close the page after submission. Notification mails will also be sent to the email address when the analysis process is finished (either success or failure).



Basic Information

please provide your basic information.

Email address:

Please re-enter your Email for comfirmation.

Re-enter Email address:

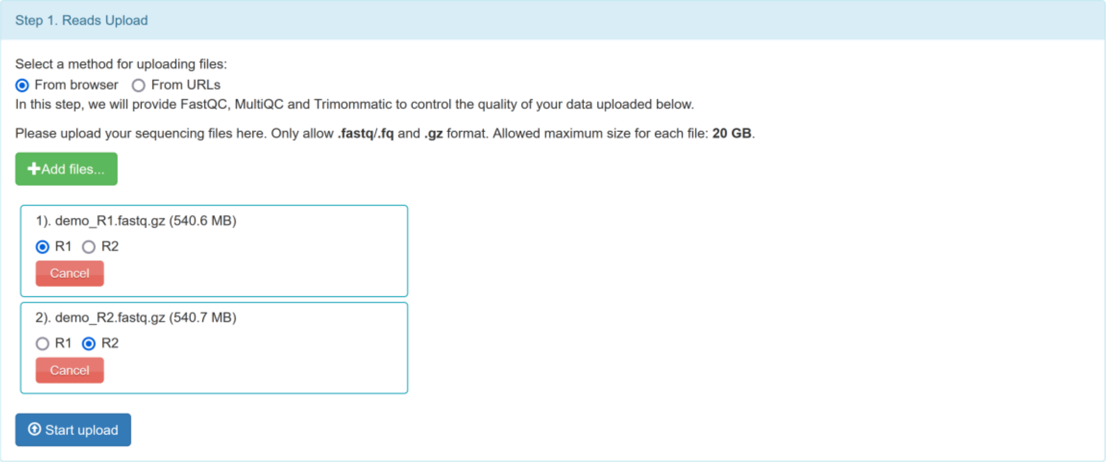
Fig.2. MiDSystem will send a unique link for your task according to the given email address.

Task Submission

Data upload

Data upload is required for all the pipelines. The R1/R2 and/or long-read files should be uploaded separately in **.fastq/.fq**, and **.gz** formats. The maximum size for each file is **20 GB**. Users can choose to upload through their browser (Fig. 3A) or with URLs (e.g. google drive links)(Fig. 3B). The reference-guided pipeline will require an extra reference file to upload in **.fasta/.fa**, **.fna**, and **.gz** formats.

(A)



Step 1. Reads Upload

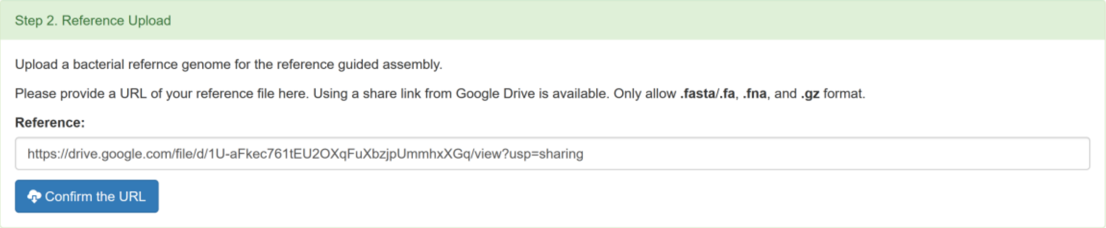
Select a method for uploading files:
● From browser ○ From URLs
In this step, we will provide FastQC, MultiQC and Trimommatic to control the quality of your data uploaded below.

Please upload your sequencing files here. Only allow **.fastq/.fq** and **.gz** format. Allowed maximum size for each file: **20 GB**.

1). demo_R1.fastq.gz (540.6 MB)
● R1 ○ R2

2). demo_R2.fastq.gz (540.7 MB)
○ R1 ● R2

(B)



Step 2. Reference Upload

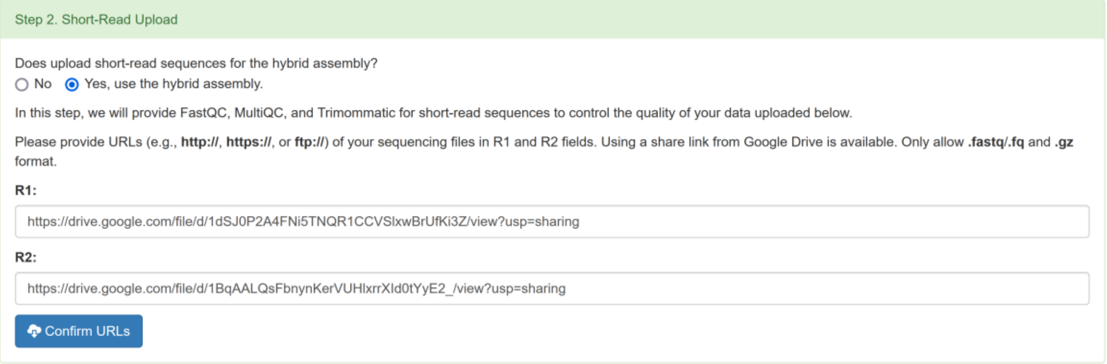
Upload a bacterial refernce genome for the reference guided assembly.

Please provide a URL of your reference file here. Using a share link from Google Drive is available. Only allow **.fasta/.fa**, **.fna**, and **.gz** format.

Reference: https://drive.google.com/file/d/1U-aFkec761tEU2OXqFuXbzjpUmmhxXGq/view?usp=sharing

Fig. 3. MiDSystem provides two upload methods for data submission. (A) Upload data from browser. (B) Upload data from an existing URL.

In Long-Read/Hybrid assembly, you can consider further uploading short-read sequences after uploading long-read sequences if available (Fig. 4). The long-read and hybrid assembly pipelines are automatically switched depending on the short-read files are provided or not.



Step 2. Short-Read Upload

Does upload short-read sequences for the hybrid assembly?

○ No ● Yes, use the hybrid assembly.

In this step, we will provide FastQC, MultiQC, and Trimommatic for short-read sequences to control the quality of your data uploaded below.

Please provide URLs (e.g., **http://**, **https://**, or **ftp://**) of your sequencing files in R1 and R2 fields. Using a share link from Google Drive is available. Only allow **.fastq/.fq** and **.gz** format.

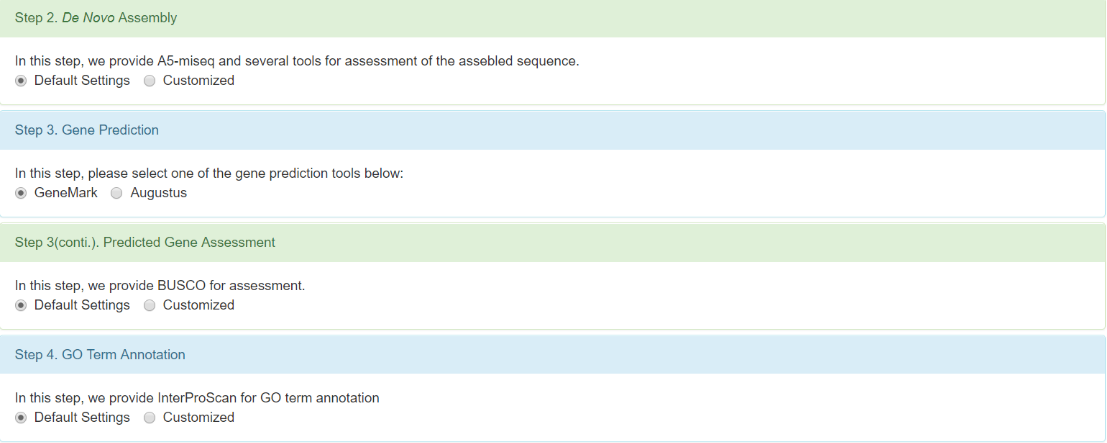
R1: https://drive.google.com/file/d/1dSJ0P2A4FNi5TNQR1CCVSlxwBrUfKi3Z/view?usp=sharing

R2: https://drive.google.com/file/d/1BqAALQsFbnynKerVUHlxrrXld0tYyE2_/view?usp=sharing

Fig. 4. Upload short-read files for conducting the hybrid assembly pipeline in Long-Read/Hybrid assembly.

Parameter settings

After file upload step, the following analysis settings are enabled. Default settings are provided to all the tools that we will run in the analysis process (Fig 5). Users can also customize the parameters based on their needs. We provide default settings, but you can use the "Customized" option to adjust the parameters (Fig 6).



Step 2. *De Novo* Assembly

In this step, we provide A5-miseq and several tools for assessment of the assebled sequence.

● Default Settings ○ Customized

Step 3. Gene Prediction

In this step, please select one of the gene prediction tools below:

● GeneMark ○ Augustus

Step 3(conti.). Predicted Gene Assessment

In this step, we provide BUSCO for assessment.

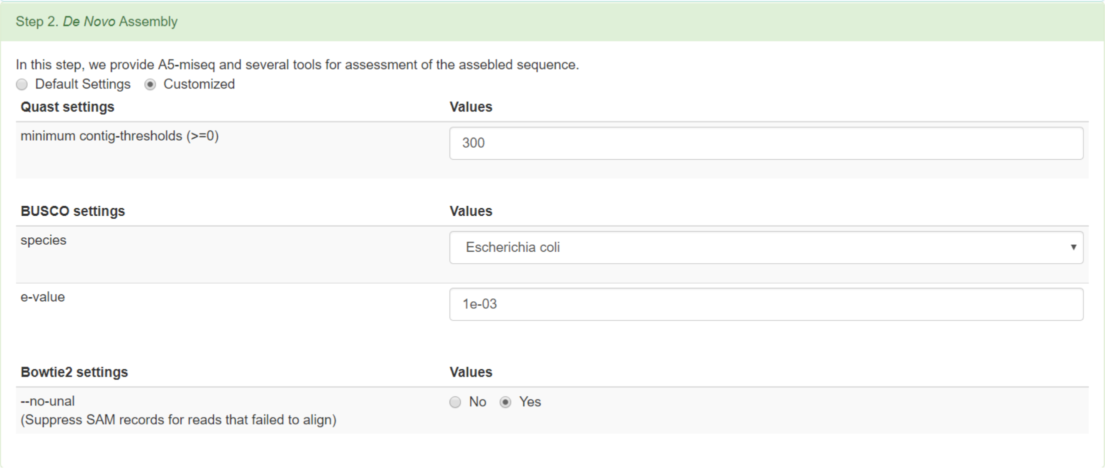
● Default Settings ○ Customized

Step 4. GO Term Annotation

In this step, we provide InterProScan for GO term annotation

● Default Settings ○ Customized

Fig. 5. Default settings of non-reference assembly.



Step 2. *De Novo* Assembly

In this step, we provide A5-miseq and several tools for assessment of the assebled sequence.

○ Default Settings ● Customized

Quast settings	Values
minimum contig-thresholds (>=0)	300

BUSCO settings	Values
species	Escherichia coli ▾
e-value	1e-03

Bowtie2 settings	Values
--no-unal (Suppress SAM records for reads that failed to align)	○ No ● Yes

Fig 6. Customized parameters for relevant tools in the *de novo* assembly step.

The parameter settings would be different depending on the selected pipeline, but the parameters that you can adjust are displayed after you choose "Customized" option. Here, we are listing parameters of assembly pipelines in MiDSystem that can be customized and their descriptions (Table 1).

Table 1. Parameters of assembly pipelines in MiDSystem.

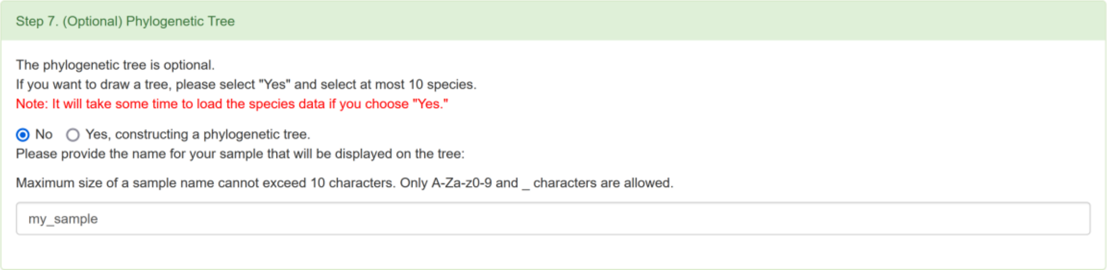
Parameter	Affected Tool	Assembly Pipeline	Default Value	Data Type	Description
mode	Unicycler	Long-read, Hybrid	Normal	{Conservative, Normal, Bold}	Conservative mode is least likely to produce a complete assembly and smaller contigs but has a very low risk of misassembly. Bold mode is most likely to produce a complete assembly and longest contigs but carries greater risk of misassembly. Normal mode is intermediate regarding both completeness and misassembly risk, as well as moderate contig size.
minimum contig-thresholds	Quast	Reference-guided (for contigs), Non-reference,	300	Non-negative integers	Lower threshold for contig length.
		Reference-guided (for scaffolds), Long-read, Hybrid	1000		
species	BUSCO	Reference-guided, Non-reference,	Escherichia coli	{Escherichia coli, Staphylococcus aureus,	Name of existing Augustus species gene finding parameters.

		Long-read, Hybrid		thermoanaerobacter_tengcongensis}	
e-value	BUSCO	Reference-guided, Non-reference, Long-read, Hybrid	1e-03	Positive floats	E-value cutoff for BLAST searches.
--no-unal	Bowtie2	Reference-guided, Non-reference, Hybrid	Yes	{Yes, No}	Suppress SAM records for reads that failed to align.
species	Augustus	Reference-guided, Non-reference, Long-read, Hybrid	Escherichia coli	{Escherichia coli, Staphylococcus aureus, thermoanaerobacter_tengcongensis}	Select a training set of specific species to predict genes.
strand	Augustus	Reference-guided, Non-reference, Long-read, Hybrid	Both	{Both, Forward, Backward}	Predict genes on which strands.
e-value	BLAST	Reference-guided, Non-reference, Long-read, Hybrid	1E-5	Positive floats	The number of hits one can “expect” to see by chance when searching a database of a particular size.
file format	InterProScan	Reference-guided, Non-reference,	tsv	{tsv, xml, gff3, svg, json, html}	Select output formats of InterProScan reports.

		Long-read, Hybrid			
output file format	GraPhlAn	Metagenomics	png	{png, pdf, ps, eps, svg}	Select output formats of cladograms.
clustering threshold	CD-HIT	Metagenomics	0.97	Positive floats (0.7-1.0)	The of threshold sequence identity. This is the default cd-hit's “global sequence identity” calculated as: number of identical amino acids in alignment divided by the full length of the shorter sequence.
insert size (-X)	Bowtie	Metagenomics	800	Positive integers (250-800)	Maximum insert size for paired-end alignment.
mismatch (-v)	Bowtie	Metagenomics	3	Non-negative integers	Report end-to-end hits with <= this number of mismatches.
e-value	HMMER	Metagenomics	1E-5	Positive floats	Report sequences <= this e-value threshold in output.

Phylogenetic Tree

The phylogenetic tree step is available in all assembly pipelines of single species (Fig. 7). It is optional, select “Yes” for construction of a phylogeny with your data and other 10 species. The database will be loaded in to the page and the selections of species will be enabled. (Note: If the loading is failed, try to choose “No”, then “Yes” again to reload the data.) You can name your sample with at most 10 characters, only A-Za-z0-9 and _ are allowed. The default value is “my_sample” (i.e., the tree will display my_sample to represent your sample).



Step 7. (Optional) Phylogenetic Tree

The phylogenetic tree is optional.
If you want to draw a tree, please select "Yes" and select at most 10 species.
Note: It will take some time to load the species data if you choose "Yes."

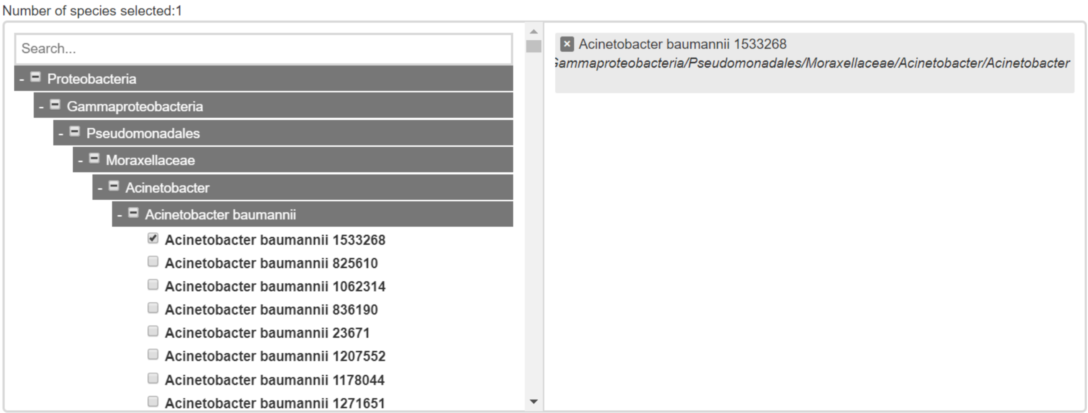
● No ○ Yes, constructing a phylogenetic tree.
Please provide the name for your sample that will be displayed on the tree:

Maximum size of a sample name cannot exceed 10 characters. Only A-Za-z0-9 and _ characters are allowed.

my_sample

Fig. 7. Setting for construction of a phylogenetic tree.

The “Search” box helps find out the species of interest. The species that you selected will be displayed on the right panel (Fig. 8). (Note: when you select more than 10 species, the oldest ones that you selected before will be removed.)



Number of species selected:1

Search...

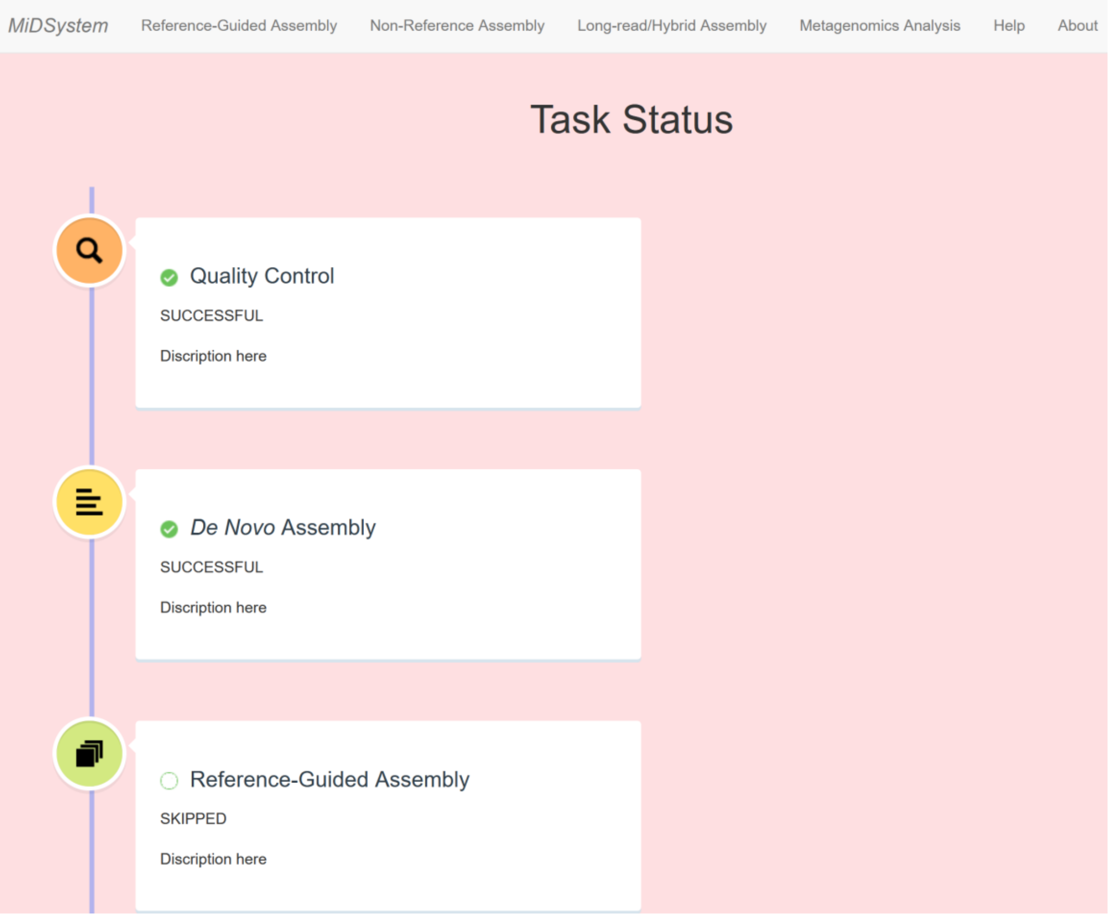
- Proteobacteria
  - Gammaproteobacteria
    - Pseudomonadales
      - Moraxellaceae
        - Acinetobacter
          - Acinetobacter baumannii
            - ☑ Acinetobacter baumannii 1533268
            - □ Acinetobacter baumannii 825610
            - □ Acinetobacter baumannii 1062314
            - □ Acinetobacter baumannii 836190
            - □ Acinetobacter baumannii 23671
            - □ Acinetobacter baumannii 1207552
            - □ Acinetobacter baumannii 1178044
            - □ Acinetobacter baumannii 1271651

☒ Acinetobacter baumannii 1533268
ammaproteobacteria/Pseudomonadales/Moraxellaceae/Acinetobacter/Acinetobacter

Fig. 8. Selection species in the MiDSystem database for construction of the phylogeny with uploaded data.

Task Status

After submission, you will receive an email with a link to help check out the status of the submitted task Fig. 9. A certain step will display “SKIPPED” when they are not necessary for the specific pipeline. After the task is completed, another email will be sent to the mail address you provided, and this page will be no longer available since it will be automatically redirect to a report page.



Task Status

✓ Quality Control

SUCCESSFUL

Discription here

✓ *De Novo* Assembly

SUCCESSFUL

Discription here

○ Reference-Guided Assembly

SKIPPED

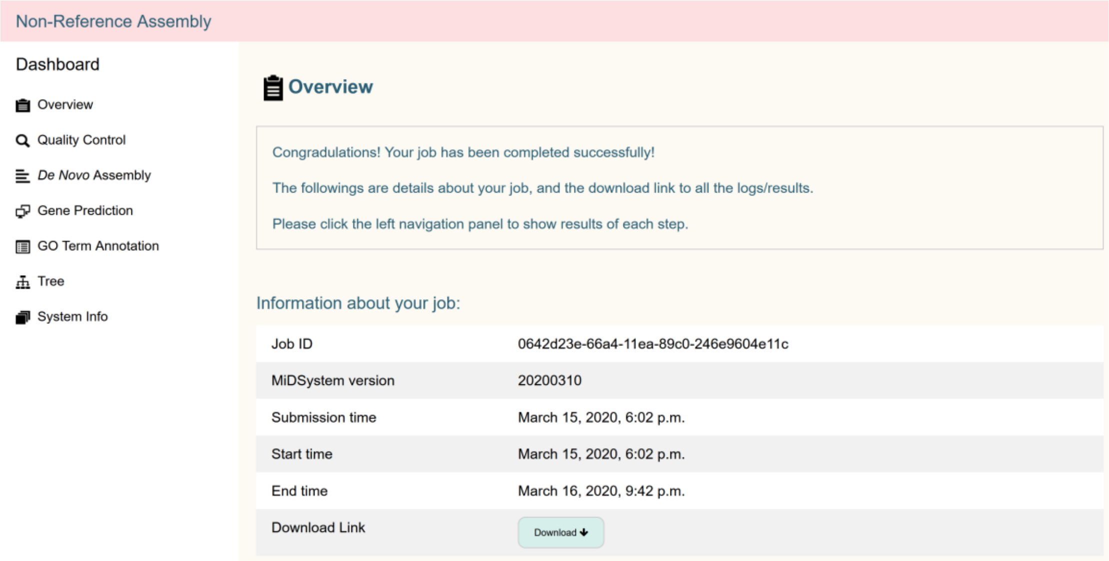
Discription here

Fig. 9. Screenshot of the task status.

Result Pages

Overview

The result page is shown in Fig. 10. The name of performed pipeline displays at the top left corner. The task information and download link of full analysis reports are provided for users. The left panel can be used to navigate through summary reports of tools. To make complicated reports that can be interpreted, all outputs in MiDSystem are summarized as tables, pie charts, bar charts, etc.



Overview

Congradulations! Your job has been completed successfully!

The followings are details about your job, and the download link to all the logs/results.

Please click the left navigation panel to show results of each step.

Information about your job:

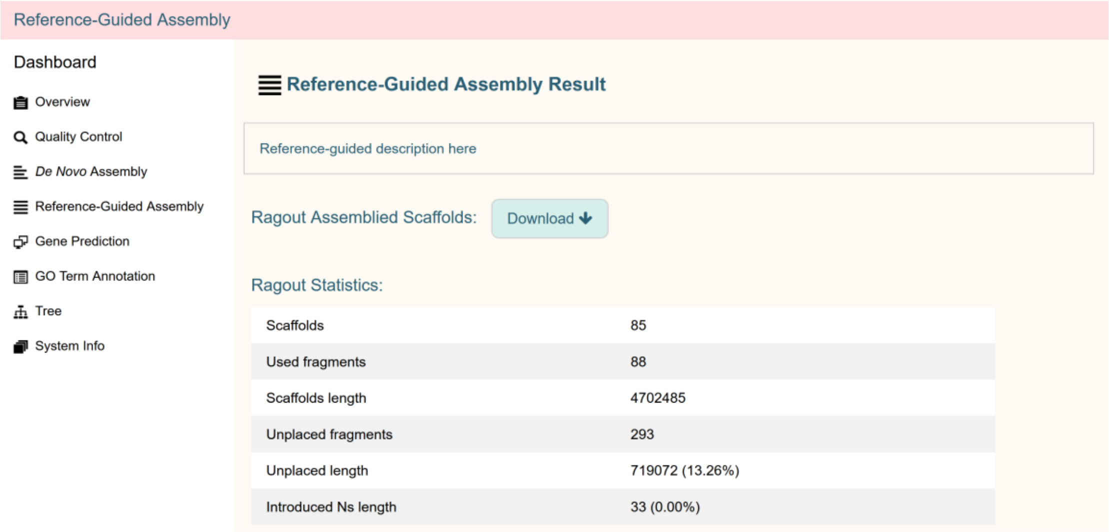
Job ID	0642d23e-66a4-11ea-89c0-246e9604e11c
MiDSystem version	20200310
Submission time	March 15, 2020, 6:02 p.m.
Start time	March 15, 2020, 6:02 p.m.
End time	March 16, 2020, 9:42 p.m.
Download Link	Download 🡇

Fig. 10. Screenshot of the result page.

Quality Control, De Novo Assembly, Gene Prediction, and GO Term Annotation

Most of the visualizations are intuitive. To sum up, the “Quality Control” page includes detailed information about the submitted and trimmed R1/R2 sequences. The “*De Nove* Assembly” and “Reference-Guided Assembly” pages provide N50 and the quality assessment reports for assembly processes. Gene prediction results are also provided with visualized assessment reports. For functional annotation, the GO term annotation bar chart is displayed corresponding to three categories.

The dashboard on the left panel will be modified based on the pipeline selected. For example, the Reference-Guided Assembly Result provides summary statistics about the assembled genome after construction of longer scaffolds from short contigs (Fig. 11), but the page link does not show on the dashboard of other pipelines. Some original reports analyzed by specific tools can be directly downloaded via provided links.



Reference-Guided Assembly Result

Reference-guided description here

Ragout Assemblied Scaffolds:

Ragout Statistics:

Scaffolds	85
Used fragments	88
Scaffolds length	4702485
Unplaced fragments	293
Unplaced length	719072 (13.26%)
Introduced Ns length	33 (0.00%)

Fig. 11. The Reference-Guided Assembly Result page shows the summary statistics of scaffolds produced by Ragout. The assembled genome can be downloaded by clicking the download button.

Three categories of the GO term annotation are displayed as bar charts in different colors (Fig. 12), i.e., biological process (green), cellular component (orange), and molecular functions (blue). The GO term bar charts provide for the *de novo* assembly pipelines of single species are based on frequency plots (i.e., how many times the specific GO term exists). However, in the metagenomics pipeline, the GO term bar charts are based on the z score. Only the top-ranked GO terms, either based on frequency or z score, for each category are displayed on website pages. Detail about z score calculation is provided in the Method section of the manuscript.

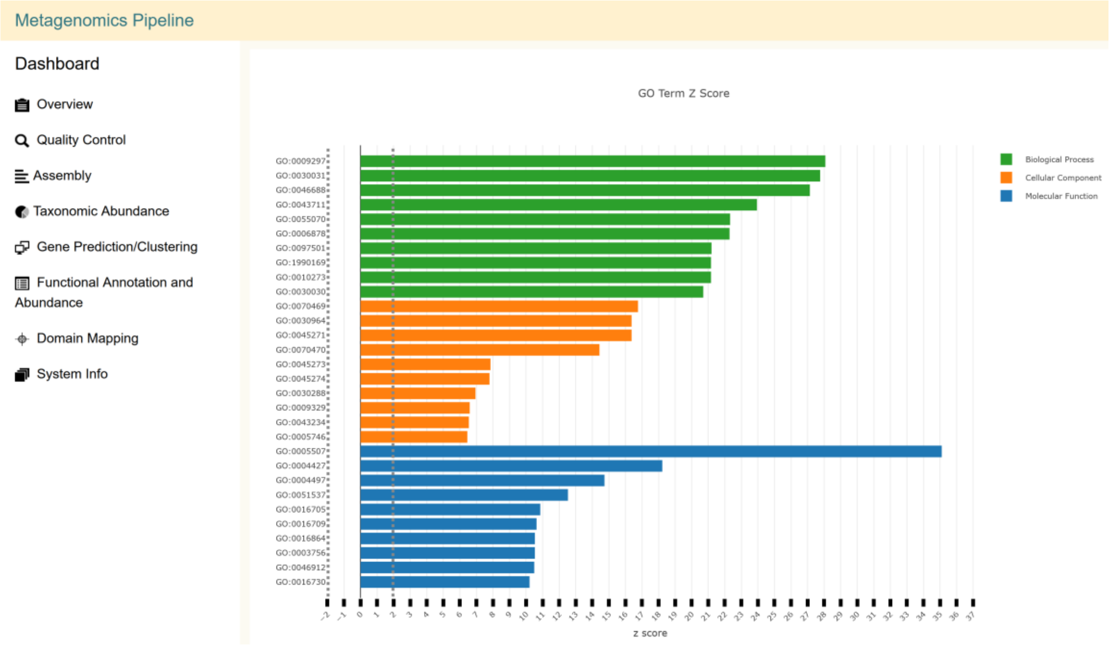
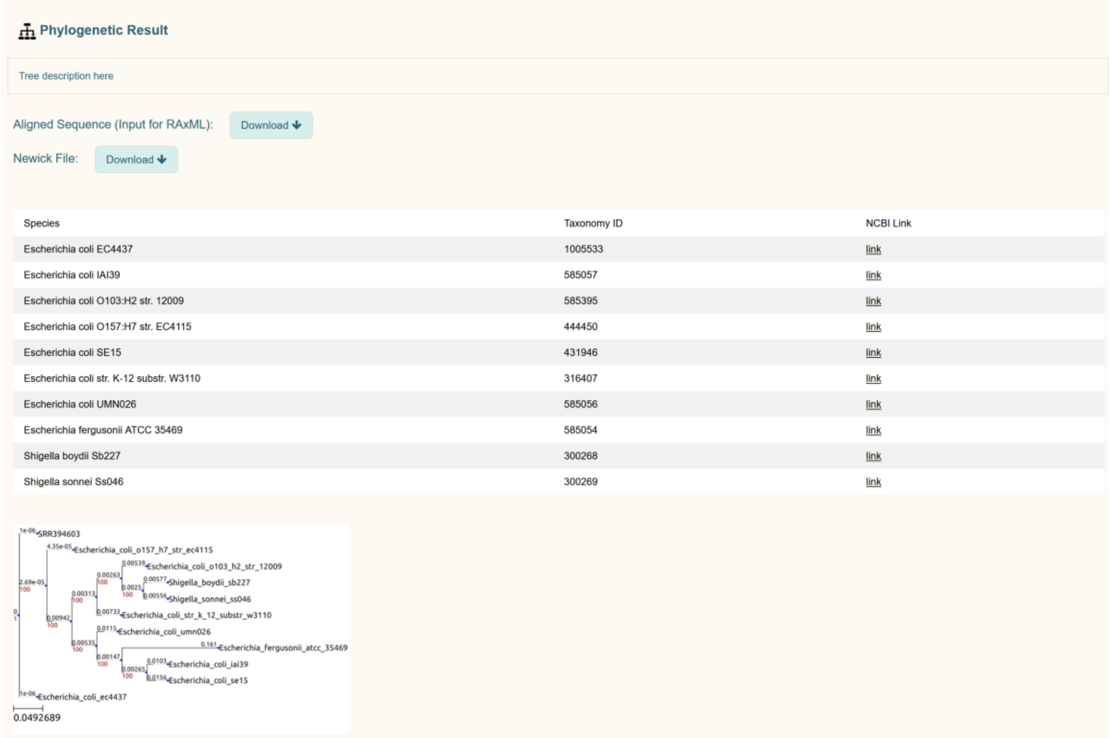


Fig. 13. The bar charts show z scores of significant GO terms in the metagenomics pipeline.

Phylogenetic Tree

Phylogenetic tree construction is only provided for the single species pipelines. A result of the *E. coli* EC4437 strain sample is provided in the manuscript case study. NCBI external links for each species selected are available on the result page (Fig. 14).



Aligned Sequence (Input for RAxML): Download

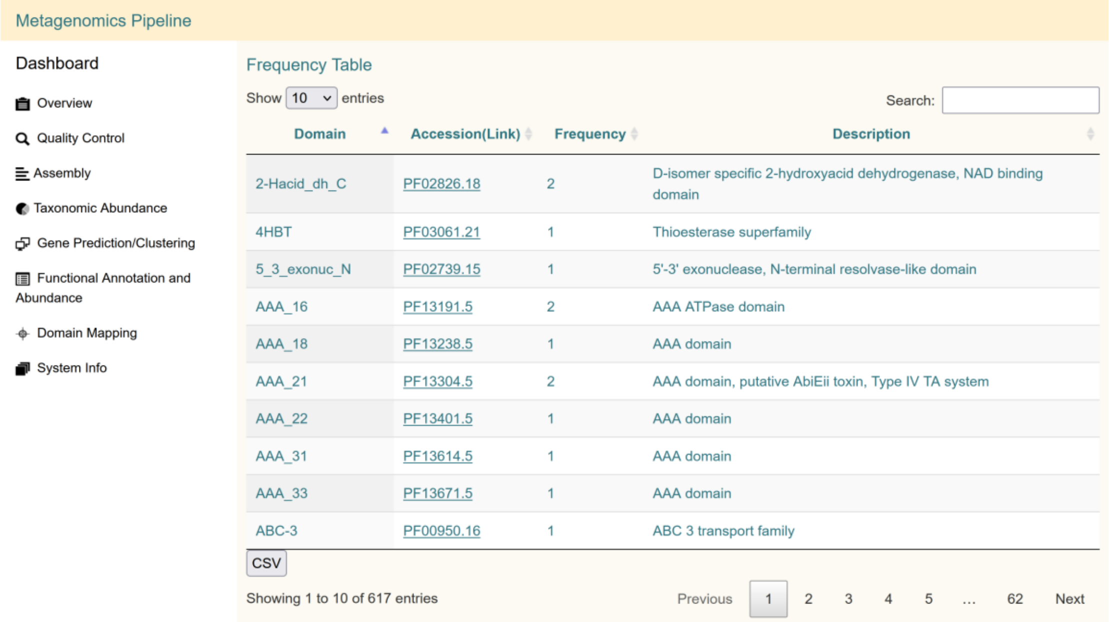
Newick File: Download

Species	Taxonomy ID	NCBI Link
Escherichia coli EC4437	1005533	link
Escherichia coli IAI39	585057	link
Escherichia coli O103:H2 str. 12009	585395	link
Escherichia coli O157:H7 str. EC4115	444450	link
Escherichia coli SE15	431946	link
Escherichia coli str. K-12 substr. W3110	316407	link
Escherichia coli UMN026	585056	link
Escherichia fergusonii ATCC 35469	585054	link
Shigella boydii Sb227	300268	link
Shigella sonnei Ss046	300269	link

Fig. 14. Screenshot of the phylogenetic result.

Result pages of the metagenomics pipeline

On the taxonomic abundance page of the metagenomics results, a cladogram, bar chart, and table are used to present the species composition in a sample. The protein domain information page is available as Fig. 15. The domain frequency and description are displayed in Frequency Table. The frequency column is the number of genes mapped to the specific domain. Accession numbers are external links to the Pfam database. This table can be downloaded in a CSV file.



Frequency Table

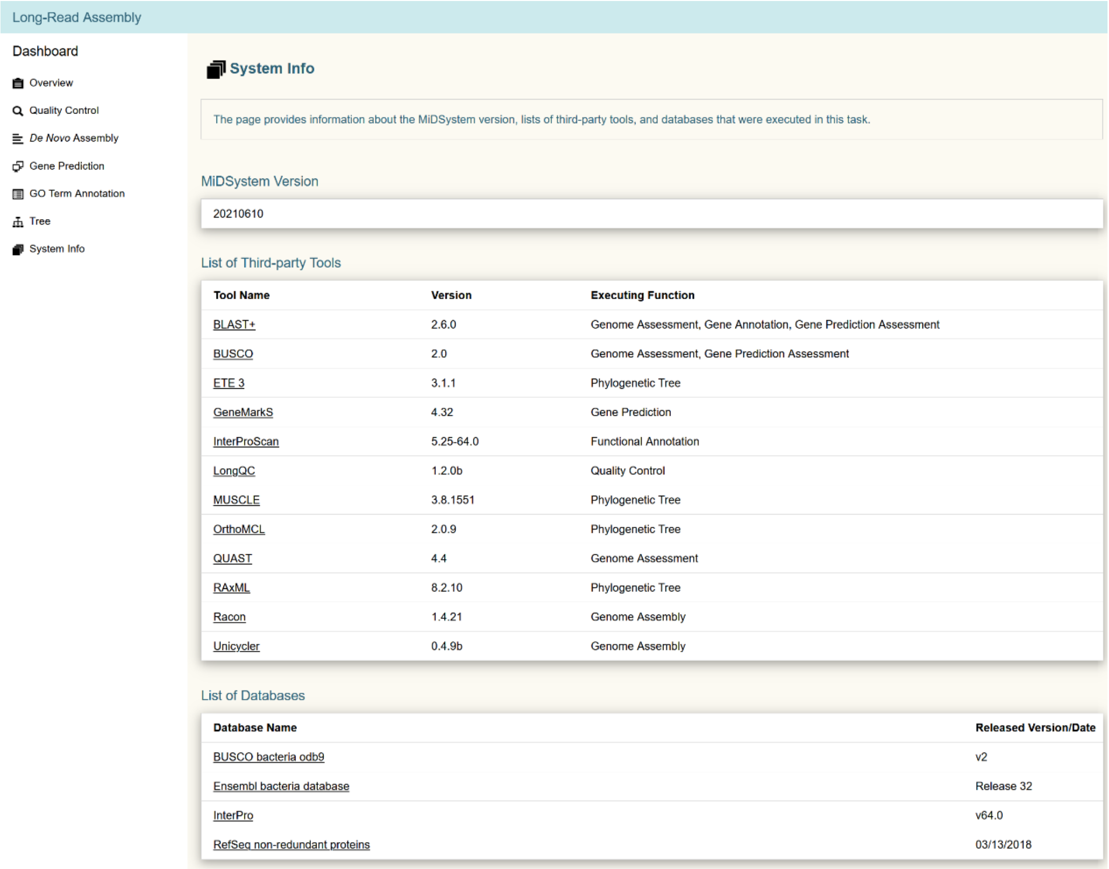
Domain	Accession(Link)	Frequency	Description
2-Hacid_dh_C	PF02826.18	2	D-isomer specific 2-hydroxyacid dehydrogenase, NAD binding domain
4HBT	PF03061.21	1	Thioesterase superfamily
5_3_exonuc_N	PF02739.15	1	5'-3' exonuclease, N-terminal resolvase-like domain
AAA_16	PF13191.5	2	AAA ATPase domain
AAA_18	PF13238.5	1	AAA domain
AAA_21	PF13304.5	2	AAA domain, putative AbiEii toxin, Type IV TA system
AAA_22	PF13401.5	1	AAA domain
AAA_31	PF13614.5	1	AAA domain
AAA_33	PF13671.5	1	AAA domain
ABC-3	PF00950.16	1	ABC 3 transport family

Showing 1 to 10 of 617 entries

Fig. 15. Frequency Table shows on the domain mapping page.

System Info

The link of System Info can be found on the dashboard in all pipelines. This page shows information about the MiDSystem version and collected all third-party tools and databases with their versions and executing functions that were conducted in the current pipeline (Fig. 16). This information would help the reproducibility of analysis results from MiDSystem. The names of tools and databases on the tables contain external links to their official websites.



System Info

The page provides information about the MiDSystem version, lists of third-party tools, and databases that were executed in this task.

MiDSystem Version

20210610

List of Third-party Tools

Tool Name	Version	Executing Function
BLAST+	2.6.0	Genome Assessment, Gene Annotation, Gene Prediction Assessment
BUSCO	2.0	Genome Assessment, Gene Prediction Assessment
ETE 3	3.1.1	Phylogenetic Tree
GeneMarkS	4.32	Gene Prediction
InterProScan	5.25-64.0	Functional Annotation
LongQC	1.2.0b	Quality Control
MUSCLE	3.8.1551	Phylogenetic Tree
OrthoMCL	2.0.9	Phylogenetic Tree
QUAST	4.4	Genome Assessment
RAxML	8.2.10	Phylogenetic Tree
Racon	1.4.21	Genome Assembly
Unicycler	0.4.9b	Genome Assembly

List of Databases

Database Name	Released Version/Date
BUSCO bacteria odb9	v2
Ensembl bacteria database	Release 32
InterPro	v64.0
RefSeq non-redundant proteins	03/13/2018

Fig. 16. Screenshot of the System Info page.