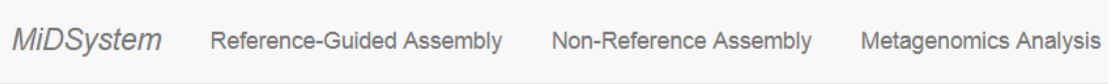
MiDSystem Quick Start

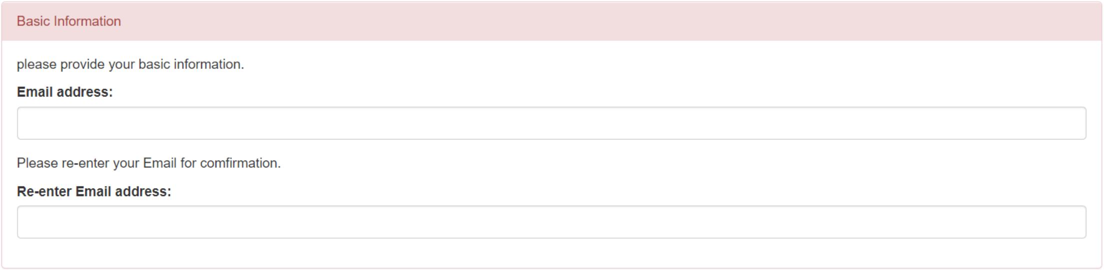
General Instructions:

The system is separated into three parts: reference-guided genome assembly, non-reference-guided assembly, and metagenomics analysis.
Reference-guided genome assembly and non-reference-guided assembly are for single microbial species. If you already have the reference genome for your sequence, we suggest you to run the reference-guided pipeline. Trying to run metagenomics samples on these two pipelines, or single species data on metagenomics pipeline will fail the analysis process since they are designed for different purpose.

The panel that allow you to choose the pipeline that you want 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 (the email address), so that we can send the unique links for you to check the status or final report of the submitted task. The unique link for your task is generated for security reason. With correct email address and the links, you can come back to check your data any time, so feel free to close the pages after submission. Notification mails will also be sent to the email address you provided when the analysis process is finished (either successful or fail.)



Basic Information

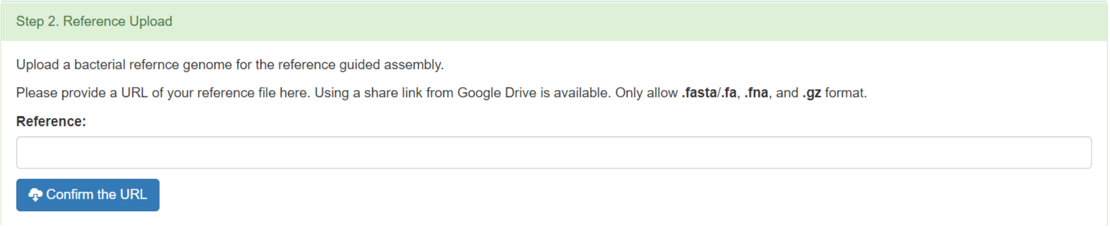
please provide your basic information.

Email address:

Please re-enter your Email for comfirmation.

Re-enter Email address:

Data upload are required for all the pipelines. Now we only accept data from Illumina platform. R1 and R2 files should be uploaded separately in **.fastq/.fq** and **.gz** format. Maximum size for each file is **20 GB**. Users can choose to upload through their browser or with URLs (e.g. google drive links). Reference-guided pipeline will require an extra reference file to upload.



Step 2. Reference Upload

Upload a bacterial refernce genome for the reference guided assembly.

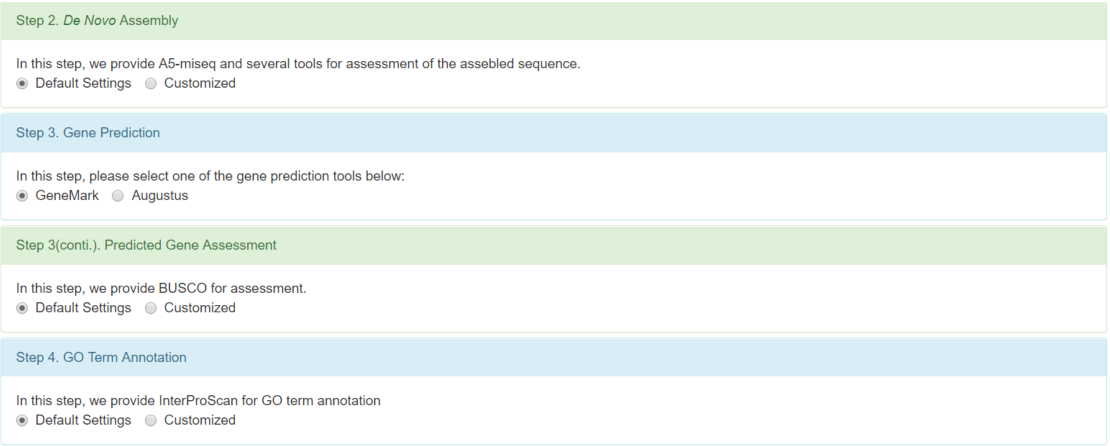
Please provide a URL of your reference file here. Using a share link from Google Drive is available. Only allow **.fasta**/**.fa**, **.fna**, and **.gz** format.

Reference:

Default settings are provided to all the tools that we will run in the analysis process. Users can also customize the options based on their needs.

Task Submission:

After file upload step, the following analysis settings are enabled. We provide default settings, but you can use the "Customized" option to adjust the parameters.
All default steps are as follow:



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

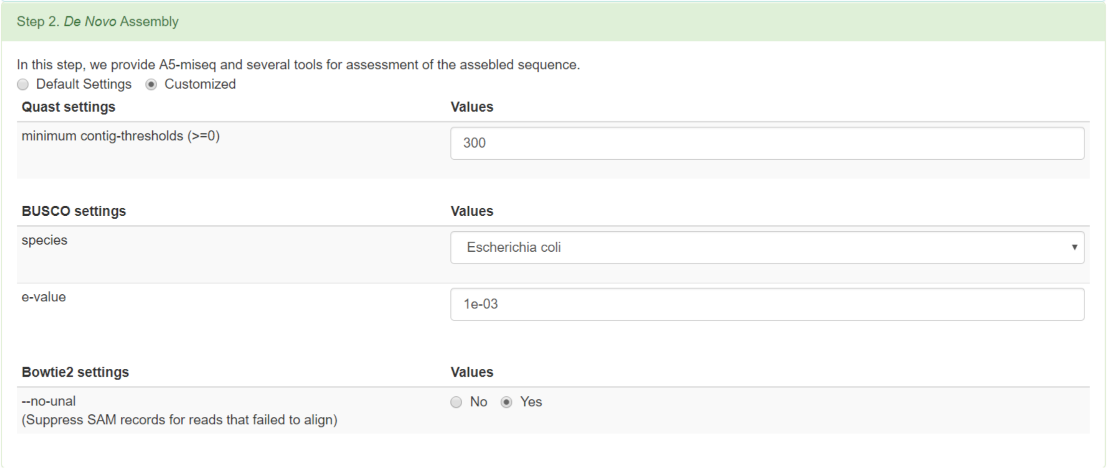
Step 3(conti.). Predicted Gene Assessment

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

Take *De Novo* Assembly for example, the parameters that you can adjust are displayed after you choose "Customized" option. For parameter setting, please refer to the tutorial of the tools. The names of the original papers are provided in the manuscript supplement section.



Step 2. *De Novo* Assembly

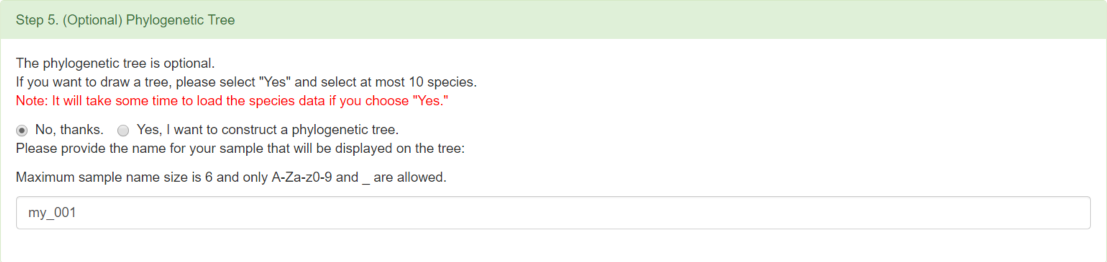
In this step, we provide A5-miseq and several tools for assessment of the assebled sequence.
○ Default Settings ● Customized

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

The Phylogenetic Tree step is optional. If you want to construct a tree with your data and other 10 species, please select "Yes." Then 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 and 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 5. (Optional) Phylogenetic Tree

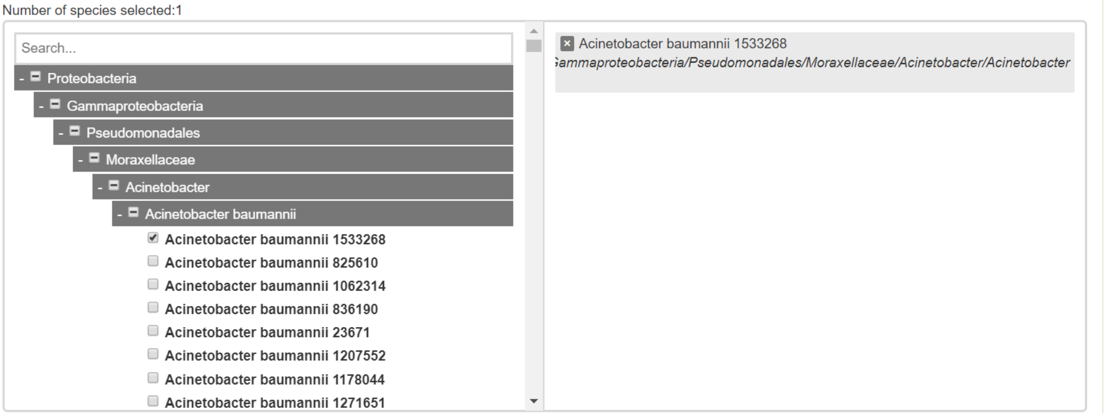
The phylogenetic tree is optional.
If you want to draw a tree, please select "Yes" and select at most 10 species.
Note: It will take some time to load the species data if you choose "Yes."

● No, thanks. ○ Yes, I want to construct a phylogenetic tree.
Please provide the name for your sample that will be displayed on the tree:

Maximum sample name size is 6 and only A-Za-z0-9 and _ are allowed.

my_001

Selection section:
Feel free to use "Search" function to find the species you want. The species that you selected will be displayed at the right panel. (Note: when you select more than 10 species, the oldest ones that you selected before will be removed.)



Number of species selected:1

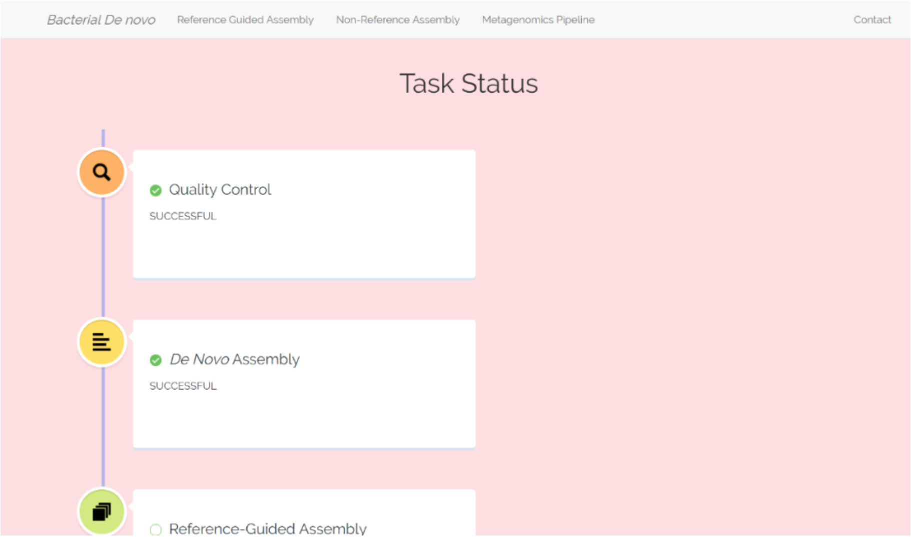
Search...

- ⊟ 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
Gammaproteobacteria/Pseudomonadales/Moraxellaceae/Acinetobacter/Acinetobacter

Click Submit, then you will receive a mail with a link to help you check the status of the submitted task as below. Some steps will display "SKIPPED" when they are not necessary for the specific pipeline. When the whole task is finished, another mail will be sent to the mail address you provided, and this page will be no longer available since it will be automatically redirect to the report pages.



Task Status

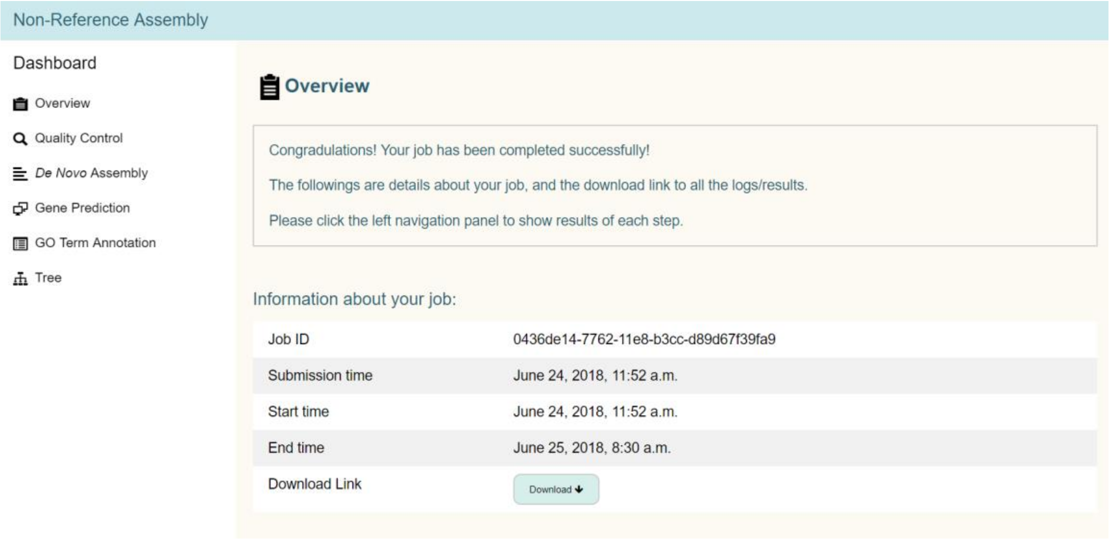
Quality Control
SUCCESSFUL

De Novo Assembly
SUCCESSFUL

Reference-Guided Assembly

Result Page:

The result page is as the screenshot below. The type of pipeline you ran will be displayed at the top left corner. Task information and download link of all the output files will be provided for users. Left panel can be used to navigate through summary reports of tools. Details about what tools are used during each step and what analysis is performed, please see the manuscript.



Overview

Congradulations! Your job has been completed successfully!

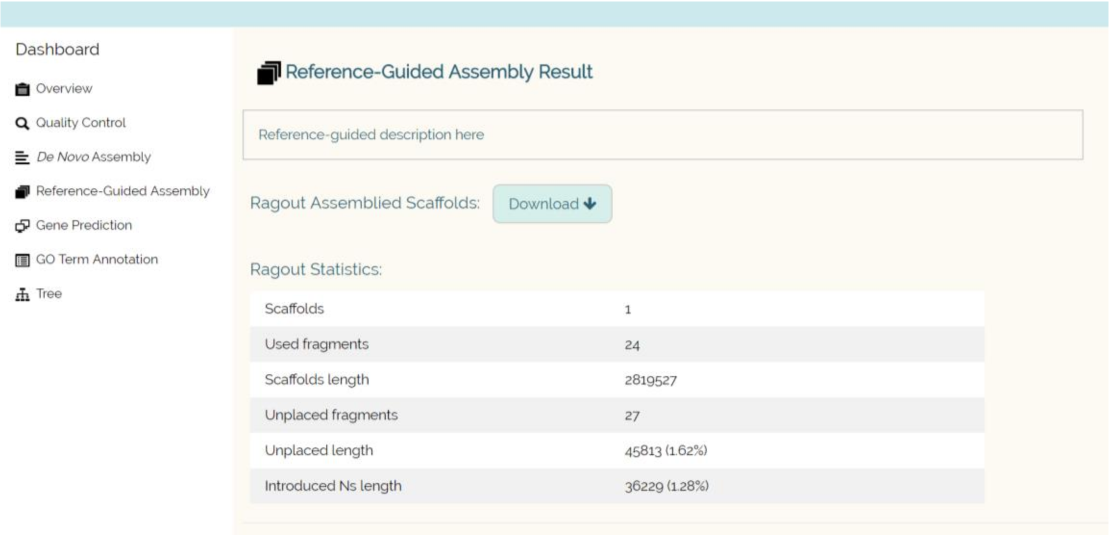
The followings are details about your job, and the download link to all the logs/results.

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

Information about your job:

Job ID	0436de14-7762-11e8-b3cc-d89d67f39fa9
Submission time	June 24, 2018, 11:52 a.m.
Start time	June 24, 2018, 11:52 a.m.
End time	June 25, 2018, 8:30 a.m.
Download Link	Download ↓

The dashboard on the left panel will be modified based on the pipeline selected. The reference-guided result page provides detail information about scaffolds after assembling short contigs from A5-miseq. The output of the tools will be display in tables, pie charts, and/or bar charts. Download links and short descriptions will also be provided to the users. Like below:



Reference-guided description here

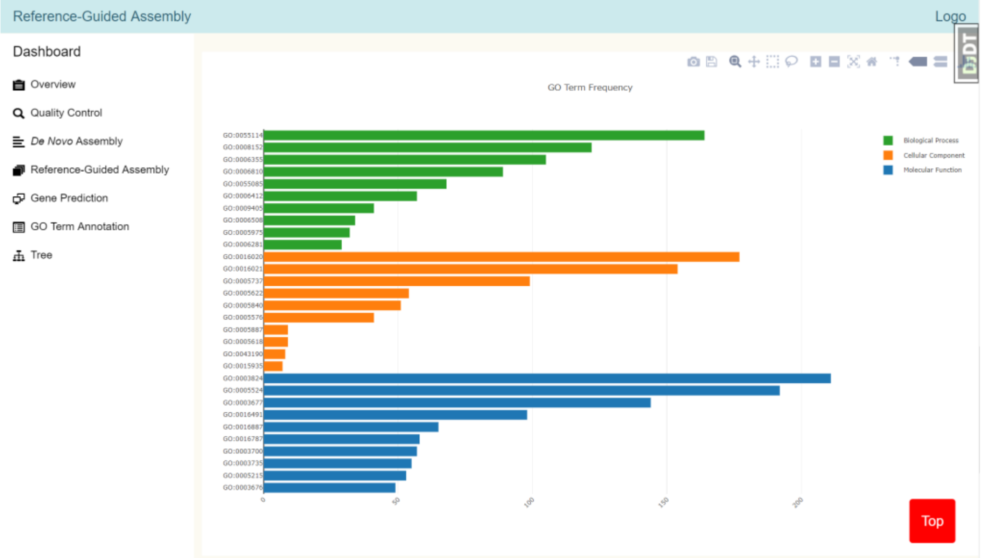
Ragout Assemblied Scaffolds: Download

Ragout Statistics:

Scaffolds	1
Used fragments	24
Scaffolds length	2819527
Unplaced fragments	27
Unplaced length	45813 (1.62%)
Introduced Ns length	36229 (1.28%)

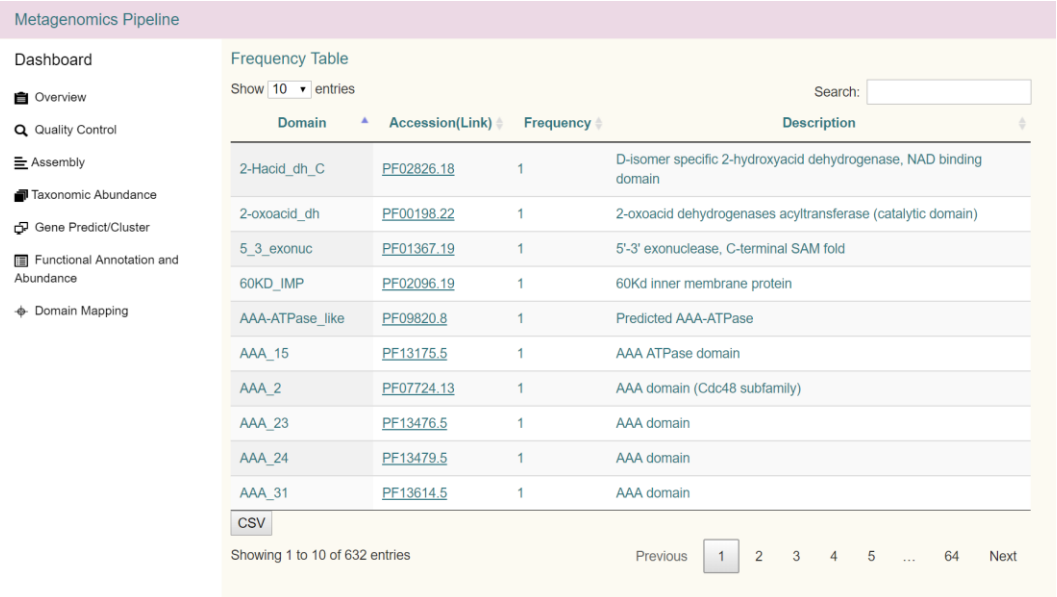
Most of the visualizations are intuitive. The “Quality Control” pages include detail information about the submitted and trimmed R1/R2 sequence. “Assembly” pages provide N50 and the quality assessment reports for assembly process. Gene prediction results are also provided with visualized assessment reports. For functional annotation, the GO term annotation bar chart is displayed corresponding to the three categories: biological process (green), cellular component (orange), and molecular functions (blue). Detail result tables can be downloaded for further research.

The GO term bar charts provided for *de novo* assembly pipelines are frequency plot (i.e. how many times the specific GO term exists.) For metagenomics pipeline, the GO term bar chart is for 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 Method section of manuscript.



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

For metagenomics pipeline, taxonomic abundance result also has a detail sample case study, including several cladograms, provided in manuscript. An additional taxonomic abundance bar chart and table are provided on website pages. The protein domain information page is available as below. Detail information about domains is displayed in the frequency table. Accession numbers are external links to Pfam database. Frequency is the number of genes mapped to the specific domain. Results can be downloaded in csv format.



Frequency Table

Domain	Accession(Link)	Frequency	Description
2-Hacid_dh_C	PF02826.18	1	D-isomer specific 2-hydroxyacid dehydrogenase, NAD binding domain
2-oxoacid_dh	PF00198.22	1	2-oxoacid dehydrogenases acyltransferase (catalytic domain)
5_3_exonuc	PF01367.19	1	5'-3' exonuclease, C-terminal SAM fold
60KD_IMP	PF02096.19	1	60Kd inner membrane protein
AAA-ATPase_like	PF09820.8	1	Predicted AAA-ATPase
AAA_15	PF13175.5	1	AAA ATPase domain
AAA_2	PF07724.13	1	AAA domain (Cdc48 subfamily)
AAA_23	PF13476.5	1	AAA domain
AAA_24	PF13479.5	1	AAA domain
AAA_31	PF13614.5	1	AAA domain

Showing 1 to 10 of 632 entries