Dr. Tom System Data Upload Procedure

Dr. Tom supports users to upload their own expression data for analysis.

This step-by-step guide shows how to upload expression data for further analysis and data mining. Expression data refers to expression matrix data, commonly obtained after alignment and quantification procedures. This guide is not for uploading original sequencing read data in FASTQ files or any other format directly obtained from sequencer.

Dr. Tom also supports users to upload DMR data for methylation analysis, though not included in this guide. Please consult your sales representative for further support.

1. Prepare expression data file to be uploaded:

An expression data file should be formatted in a 2-dimentianal table, exhibiting expression data in a “Sample Name vs. Gene ID” manner, as the example shown below.

ID	Sample_1	Sample_2	Sample_3	Sample_4	Sample_5	Sample_6
8693	6619.16	5833.86	6487.22	8363.36	5078.79	1895.82
100533467	4435.23	9811.50	8970.88	7172.33	9682.41	6750.40
109504726	141.38	8195.87	3663.18	3900.02	7942.50	8246.08
79008	6233.01	4321.49	2312.22	6103.42	8092.53	2593.47
101059918	9838.78	7165.28	5714.44	5536.26	7148.92	6679.49
150094	8389.71	5024.98	9918.34	8565.71	1773.94	7877.94
100526772	1811.06	6780.09	5695.26	8266.29	193.98	9857.68
151742	4341.98	3407.90	1639.30	8308.01	8459.63	7671.89
145781	5113.20	2465.99	439.02	6533.93	7830.16	9024.45
106865373	157.16	7089.87	4614.55	3227.41	6169.06	6638.04
116804918	5470.72	7782.15	263.75	6746.95	6989.37	5605.16
100526832	57.84	6092.46	9775.87	3709.64	5754.11	423.26
55096	8837.17	5591.61	467.75	2308.44	5335.85	6769.45

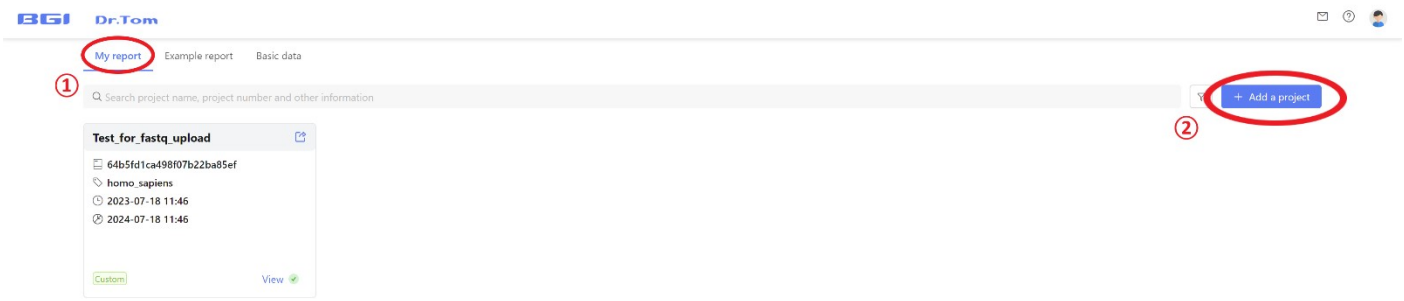
Data table can be prepared with Microsoft Excel.

Header of the first column should be “ID”. Gene ID/transcript ID or protein ID should be listed up in the following cells in the same column. From the second column, headers should be sample names and corresponding read counts/TPM/FPKM of each sample should be listed up in each column respectively. (For more detailed requirement, see Section 5)

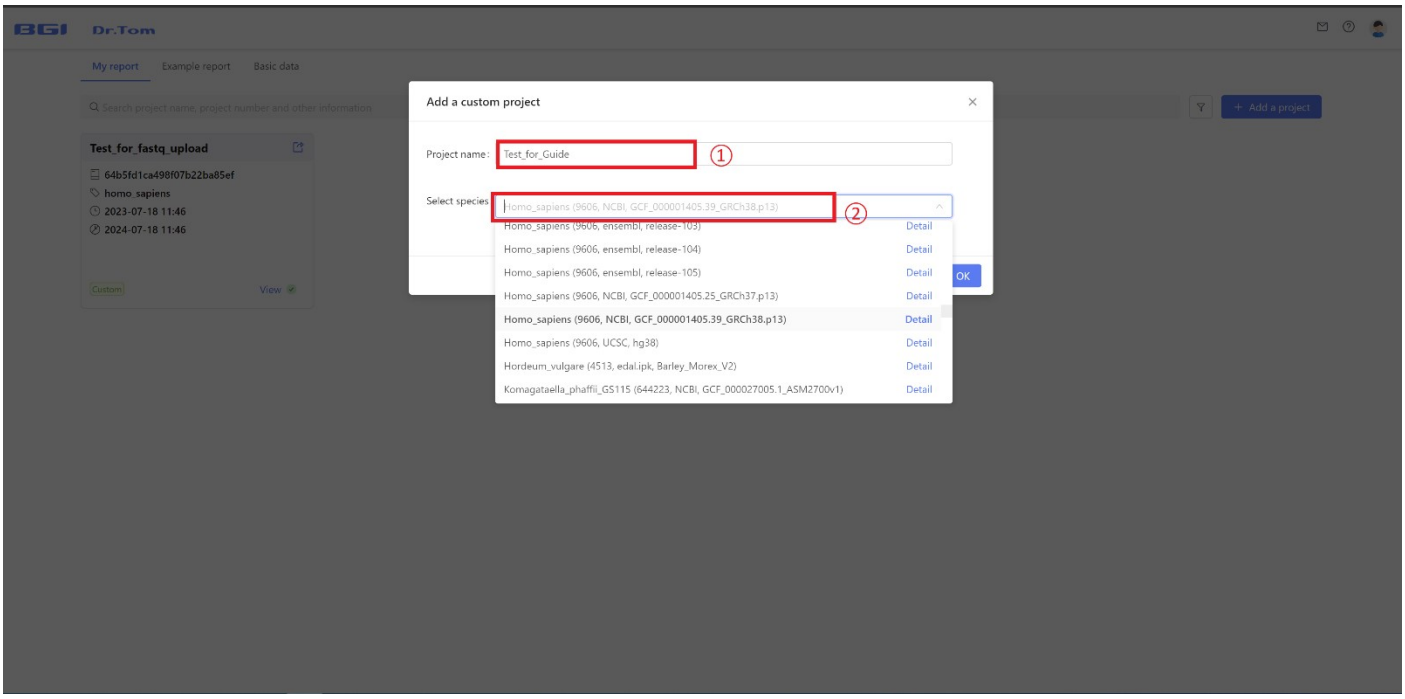
Data is suggested to be saved as a tab-delimited text file. Select "File -> Save As”, for the file type, select "Text File (Tab-Delimited) (*.txt)".

2. Upload data file

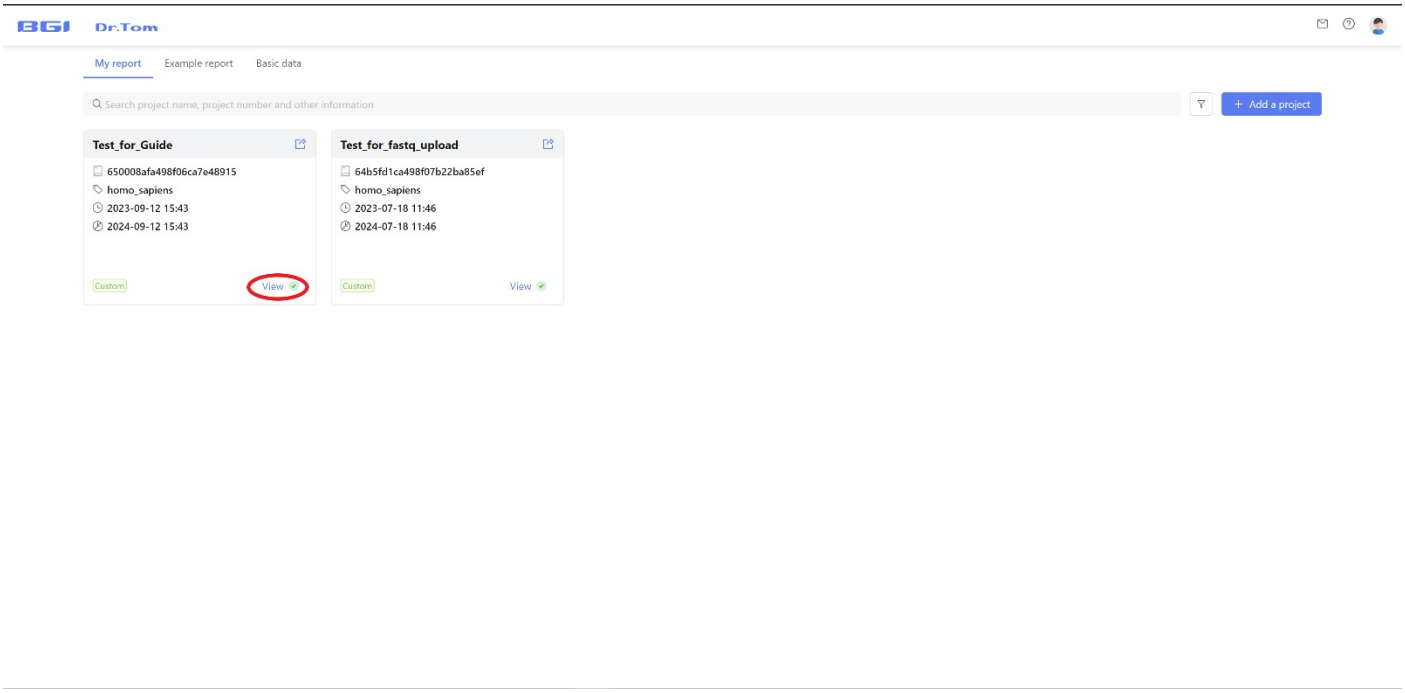
(1) Register or log in to the account. Click “+ Add a project” in the “My report” view.



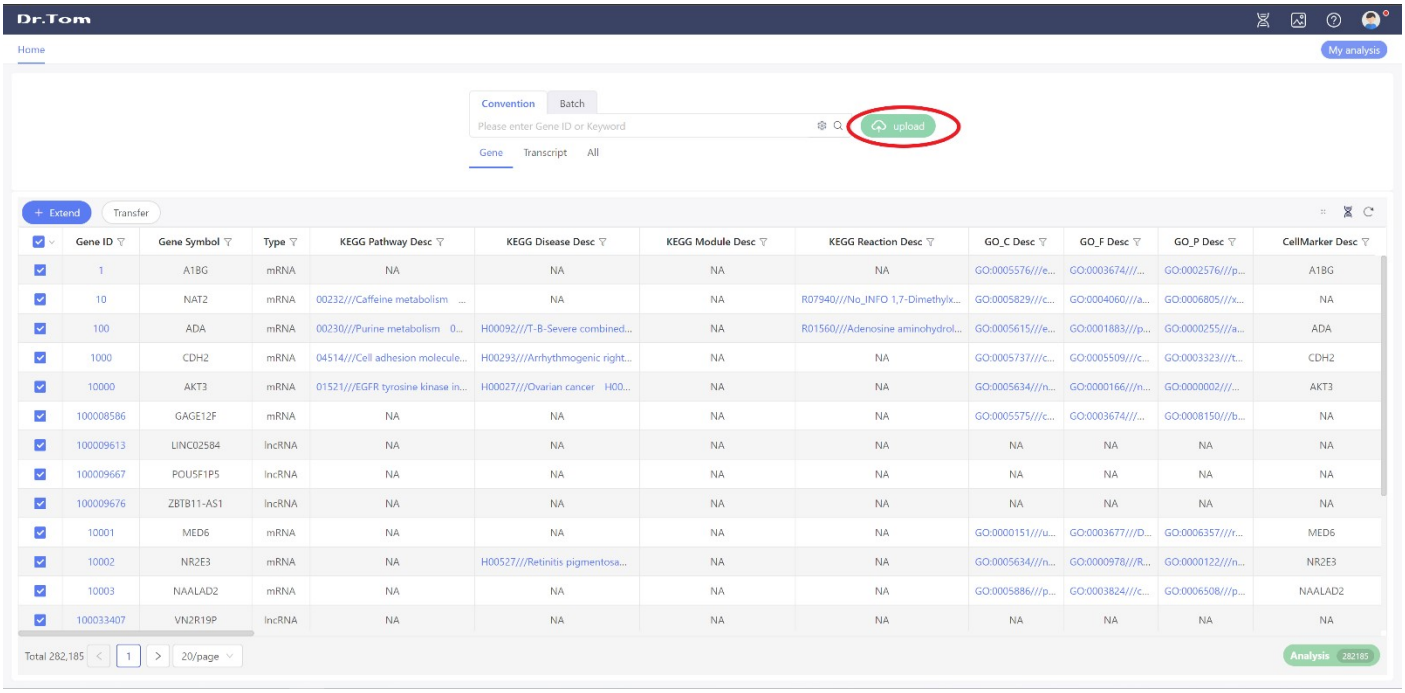
(2) Fill in “Project name”. Select reference genome in the drop-down list (Cannot be changed later! Version of reference genome must be the same as used in alignment and quantification!). Click “OK” to add.



(3) Click “View” of the new generated project to open the project page.

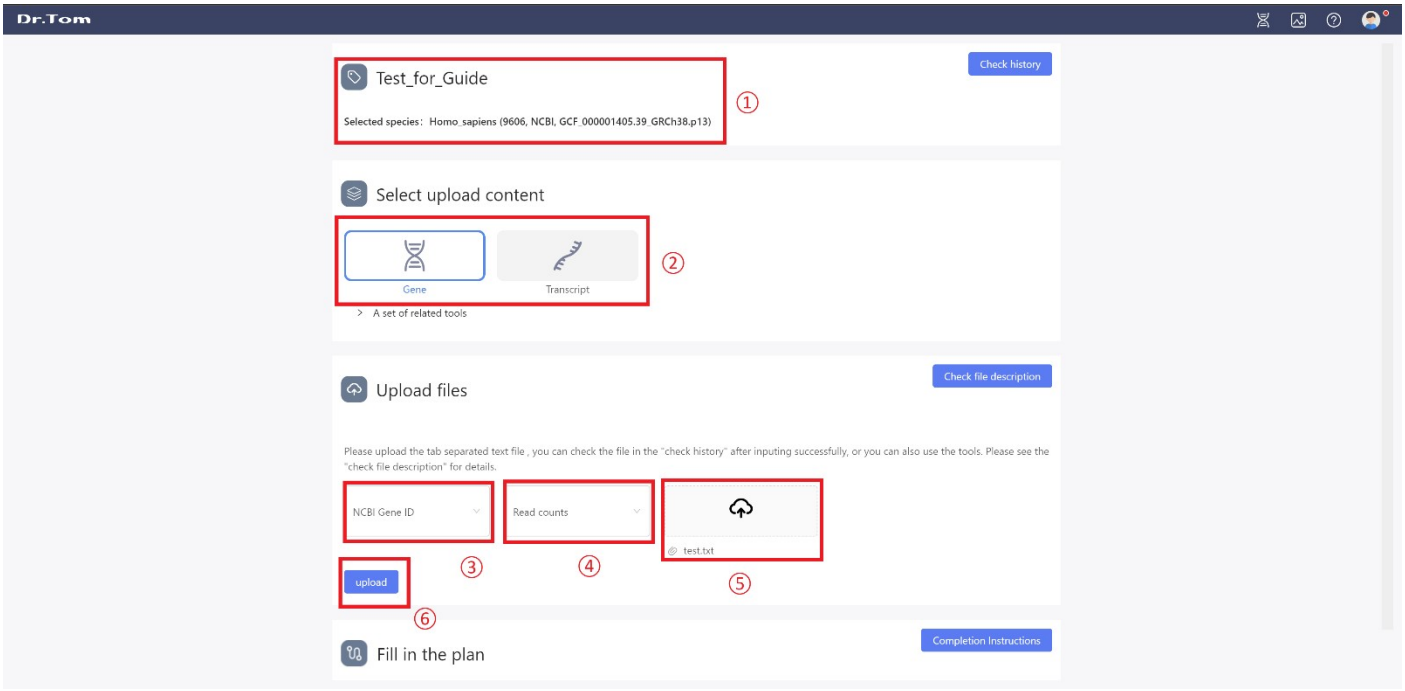


(4) Click “Upload” beside search bar to open the uploading page.

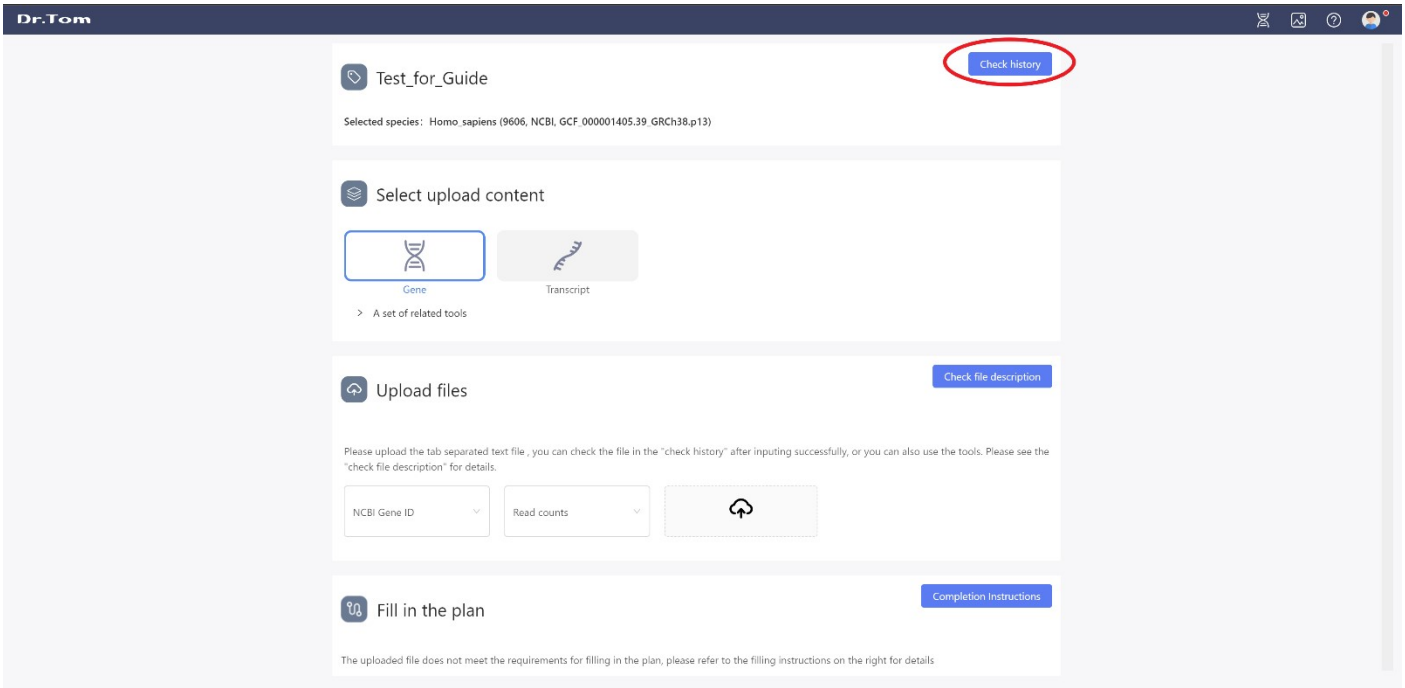


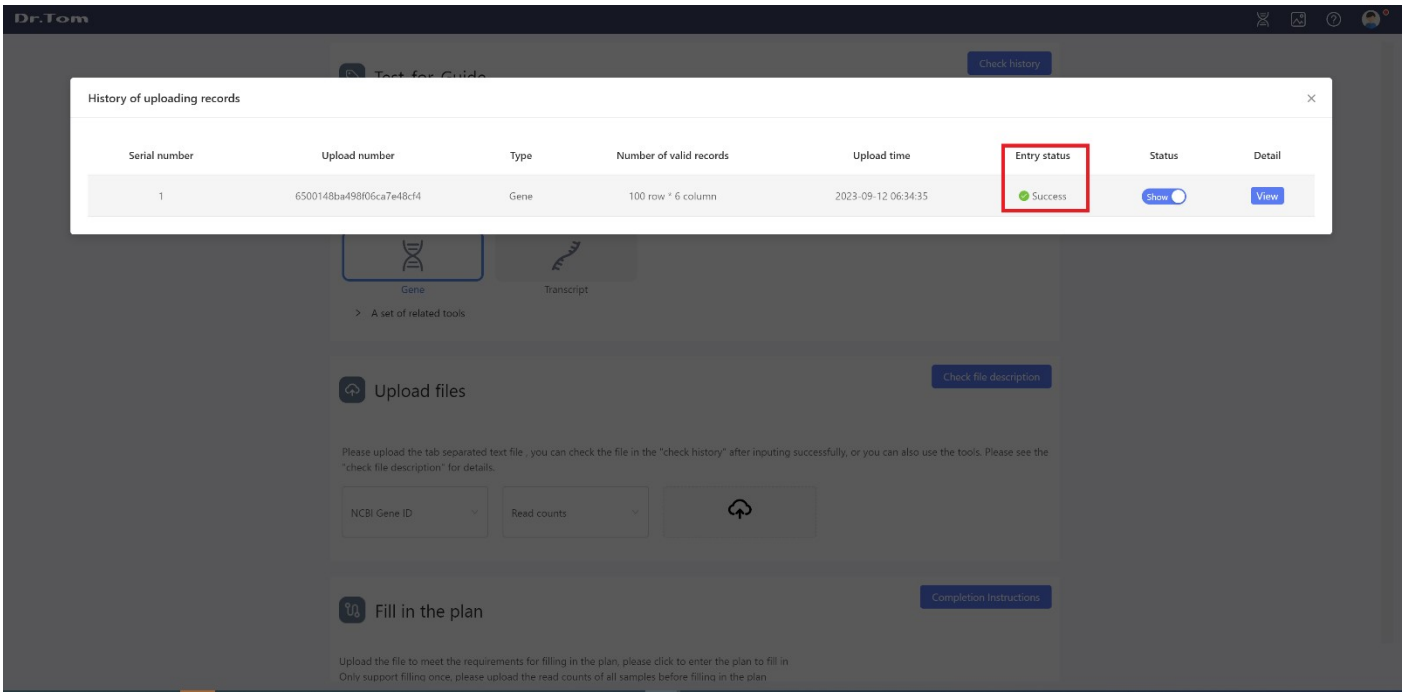
(5) Confirm if the selected species is correct. In “Select upload content” section, select “Gene” if data table is prepared in Gene ID format; or select “Transcript” if data table is prepared in Transcript ID format. In “Upload files”, choose correct ID format according to your ID, choose file type according to your expression type and then attach the data

file prepared in Section 1. Click “Upload” to upload the data.



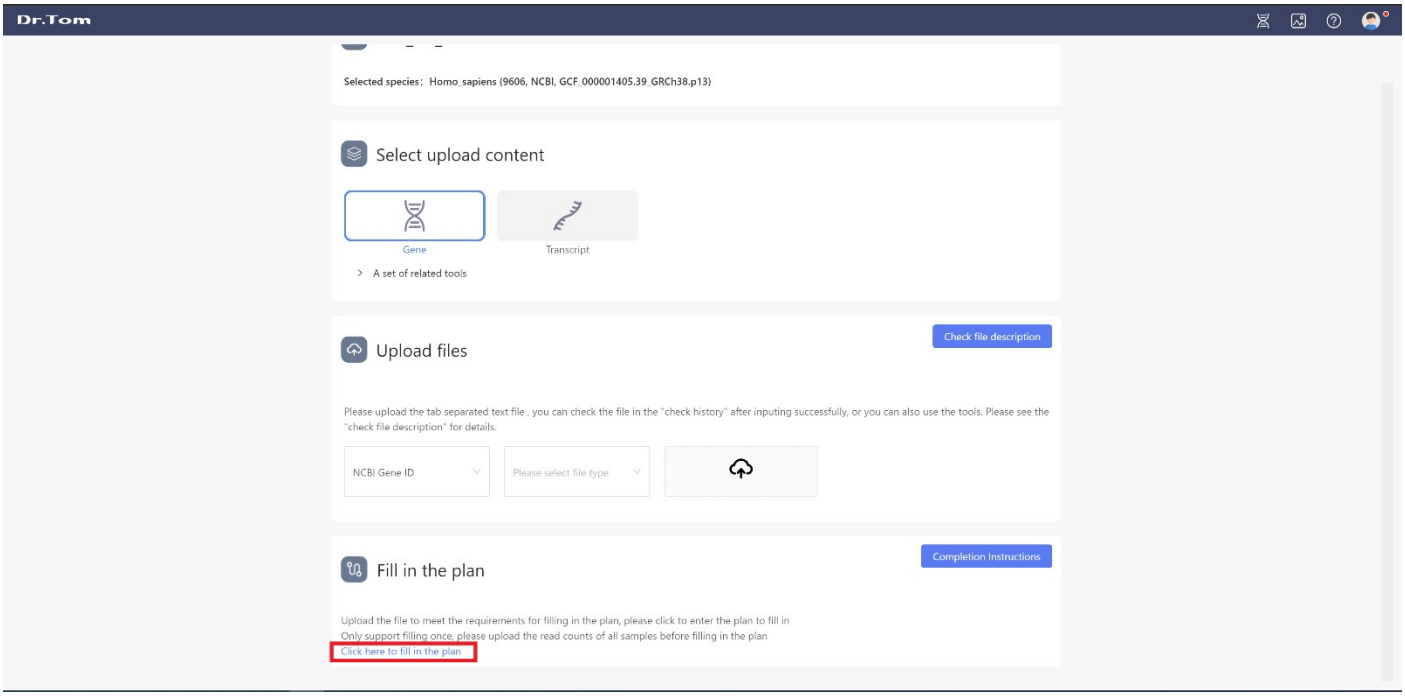
(6) When uploading finishes, click “Check History” on the top of the page. Confirm that “Entry status” is “Success”.



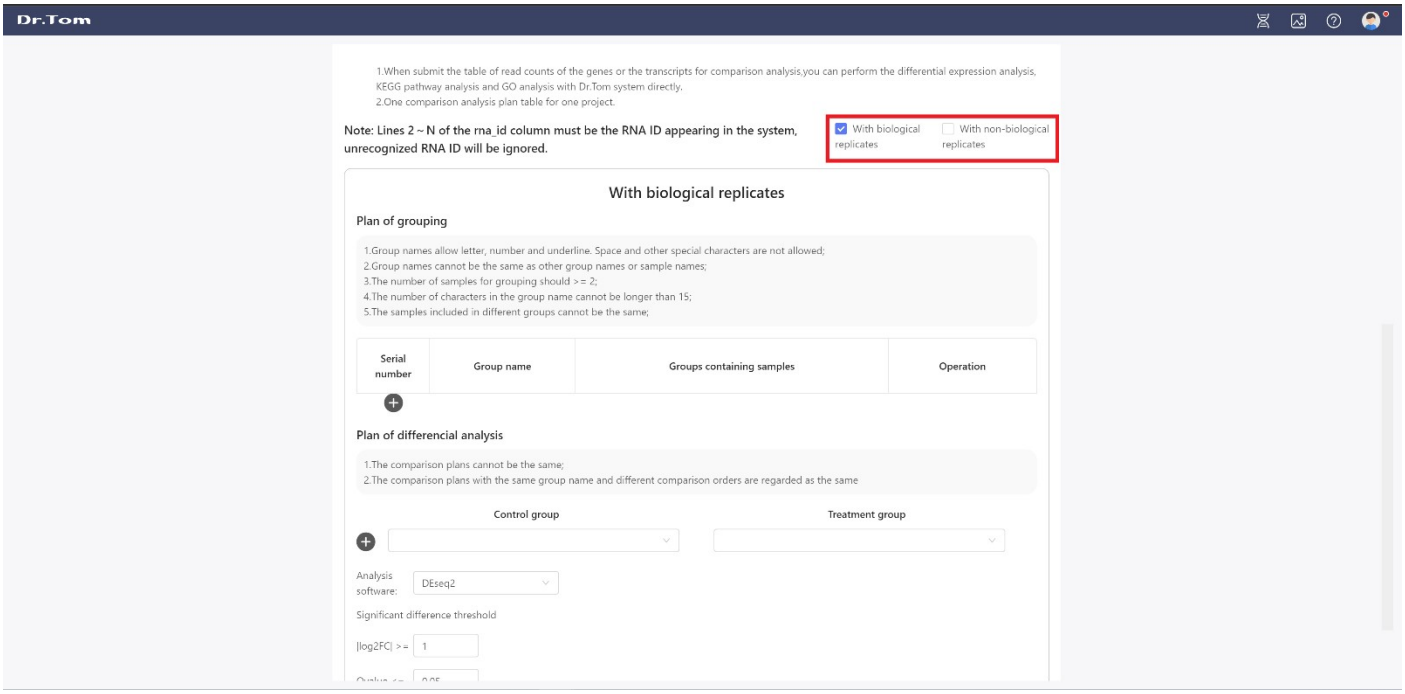


3. Fill in the analysis plan

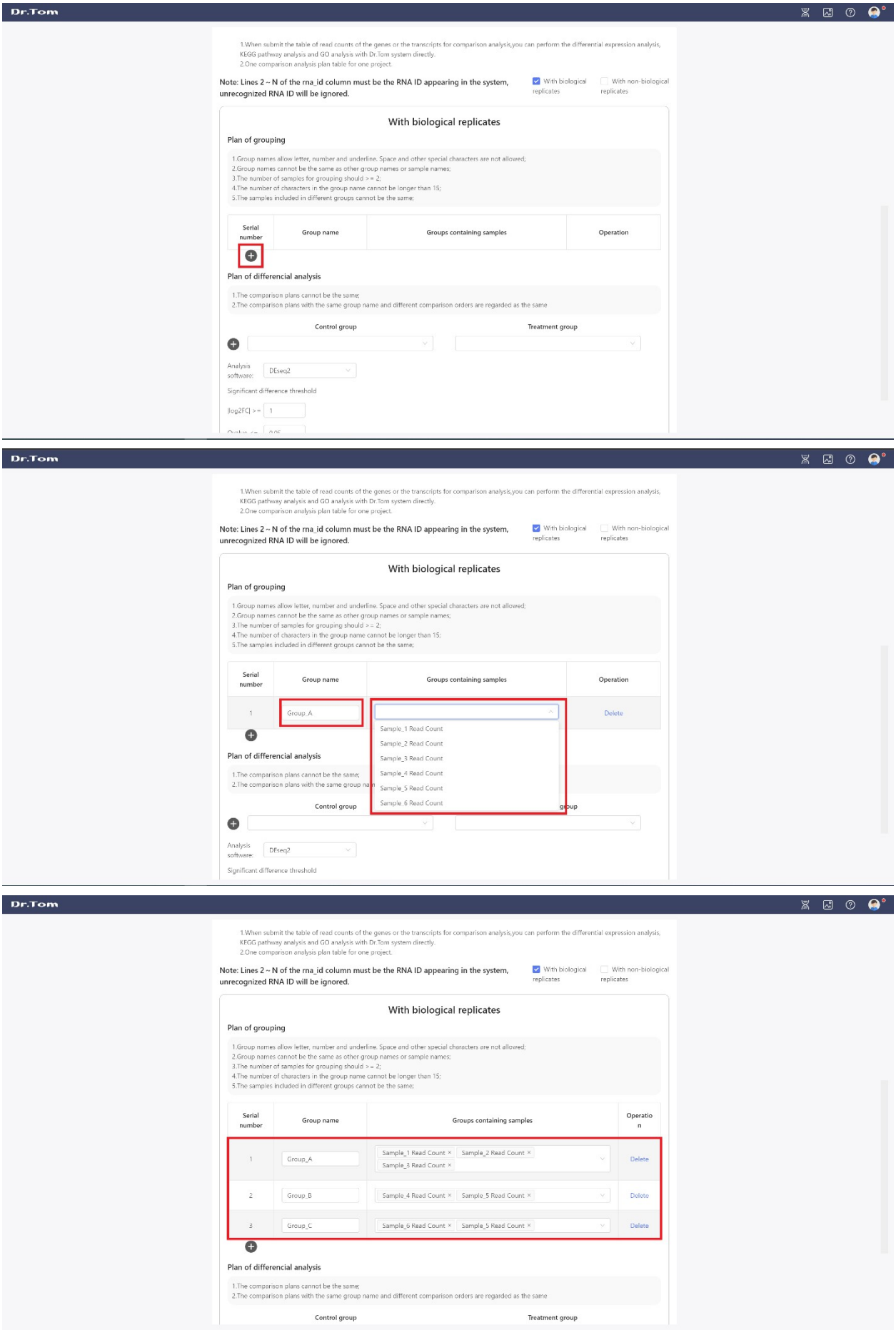
(1) Click “Click here to fill in the plan” at the bottom of the page. The analysis plan can only be filled when read count file is uploaded. If only TPM or FPKM is uploaded, the analysis plan cannot be filled.



(2) Choose “With biological replicates” if differential analysis is performed with multiple replicates in groups, or “With non-biological replicates” if analysis is performed with 1 vs 1 comparison.



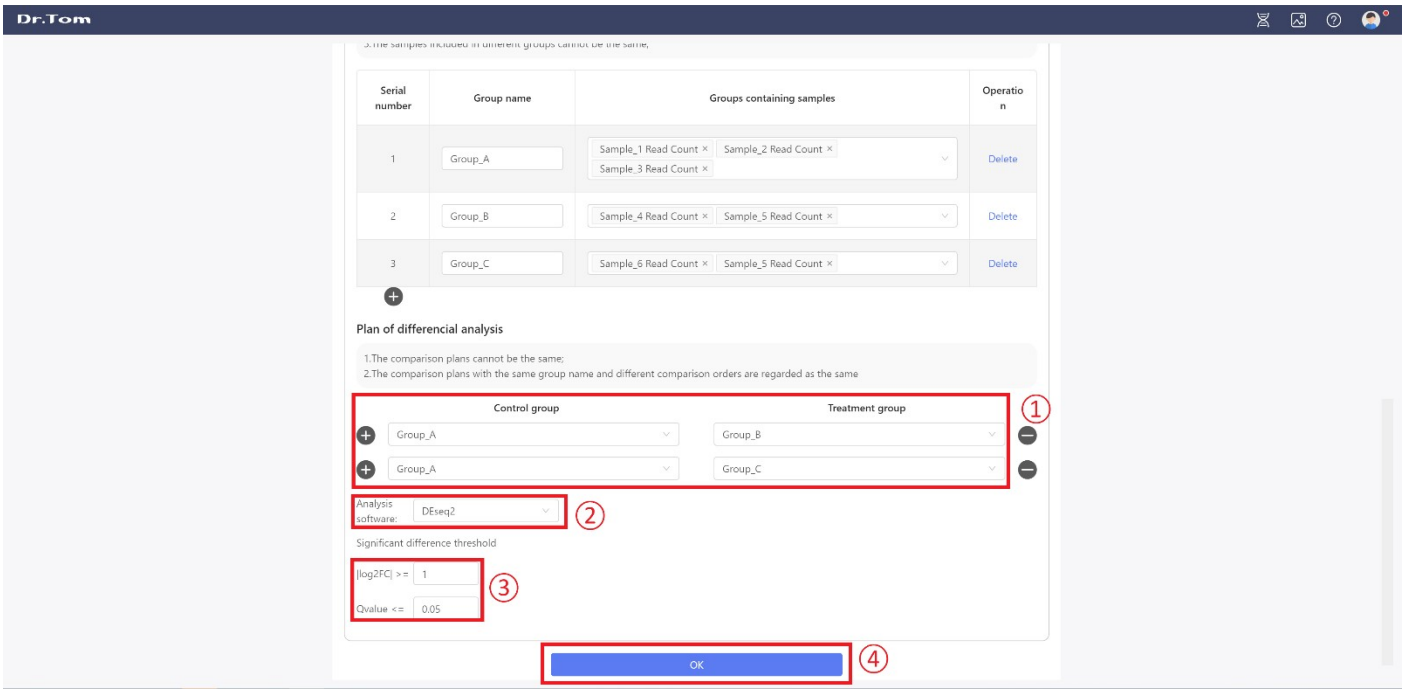
(3) Click “+” in the “Plan of grouping” to add new group; fill in “Group name” and then select corresponding samples to be put in the group; repeat until all groups are added.



(4) Select “Control group” and “Treatment Group” in “Plan of differential analysis”. Differential analysis will be performed in a format of “Treatment against Control”, i.e., Treatment/Control manner. Click “+” to add more

differential plan; choose appropriate analysis software (generally DESeq2 for comparative analysis with the sample number is >=2 in each group); set initial significant difference threshold: "|log2FC|>=" and "Q-value<="; Click "OK" to submit analysis.

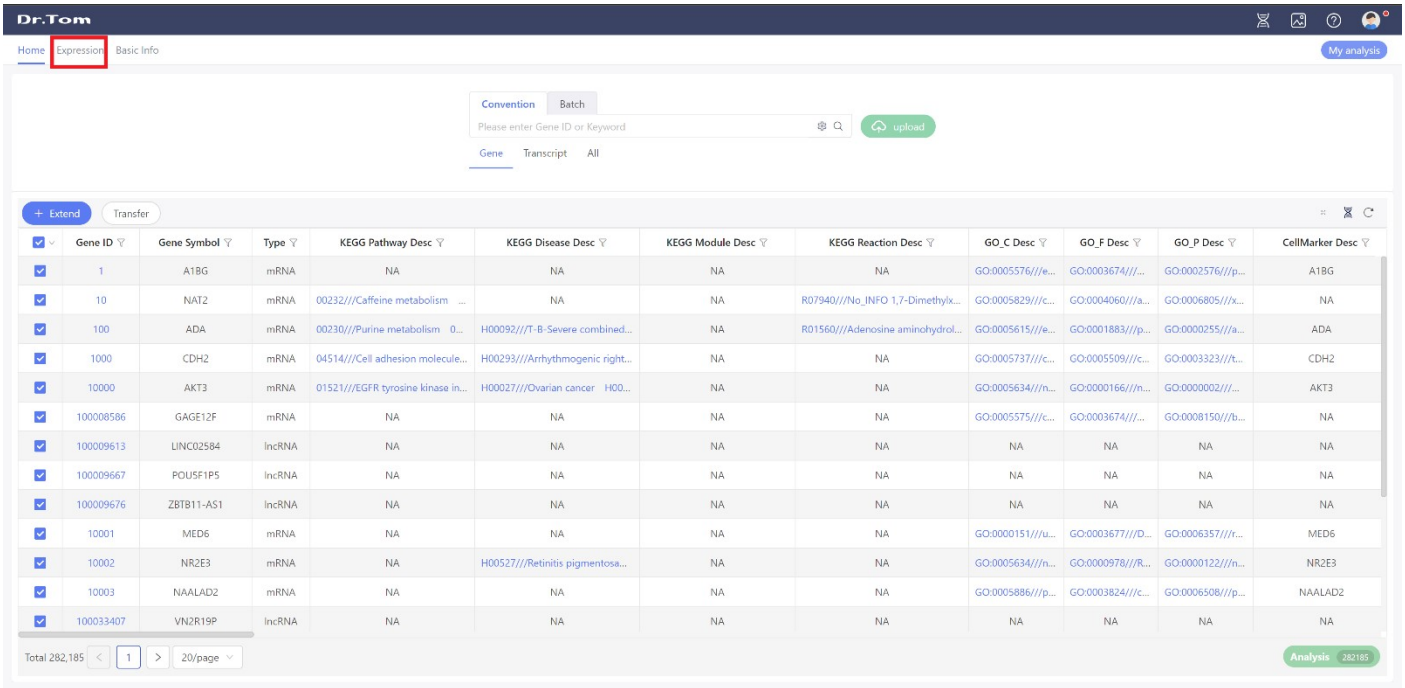
Notice: genes/transcripts with |log2FC| lower than threshold and with Q-value higher than threshold will be filtered and not included in the result. Users is suggested to set the initial threshold with a moderate threshold or re-perform differential analysis later with analysis tool panel.



4. Confirm analysis result

(1) After receiving the email titled "Dr. Tom proposal has run successfully" from bgi-drtom@bgi.com, log into Dr. Tom account.

(2) Confirm "Expression" tag is added.



(3) Start your data mining!

5. Detailed requirements

5.1 File format

File extension: Please upload a tab-delimited text file, select "File -> Save As", for the file type, select "Text File (Tab-Delimited) (*.txt)".

5.2 File size

Data size of a single upload file must be less than 20 MB.

5.3 Sample name

Sample name: It is recommended that it should not exceed 15 characters, otherwise the sample legend may be obscured and other problems may occur when plots/graphs are generated; sample name supports English letters, numbers and underscores, and does not support spaces and other special characters

5.4 ID

The ID needs to match with the selected reference genome version:

When "gene" is selected, it is mainly derived from NCBI (ID is generally purely numeric), or it may come from other public databases, and the specific source of ID can be viewed in the "Selected Species" information at the top of the data upload page. Some species support Gene Symbol and Ensembl Gene ID uploading. Check whether the current species supports it through the "ID Type" option in the upload file;

When "transcript" is selected, it mainly comes from NCBI (ID generally starts with "NM_" and "XM_"), or from other public databases, and the specific source of ID can be viewed the "Selected Species" information at the top of the data upload page;

When "protein" is selected, the ID is mainly derived from Uniprot or NCBI. If uploading data in batches in the same project, the uploaded protein ID must be consistent;

5.5 Data matrix format:

5.5.1 Genes

TPM

Header: column 1 is "ID", column 2~N is "sample name", the sample name supports English letters, numbers and underscores, and does not support spaces and other special characters;

Column 1: Gene ID;

Columns 2~N are: TPM;

FPKM

Header: column 1 is "ID", column 2~N is "sample name", the sample name supports English letters, numbers and underscores, and does not support spaces and other special characters;

Column 1: Gene ID;

Columns 2~N are: FPKM;

Read counts

Header: column 1 is "ID", column 2~N is "sample name", the sample name supports English letters, numbers and underscores, and does not support spaces and other special characters;

Column 1: Gene ID;

Columns 2~N are: read counts;

other

Header: column 1 is "ID", column 2~N is "sample name", the sample name supports English letters, numbers and underscores, and does not support spaces and other special characters;

Column 1: Gene ID;

Column 2~N: Support character type or numeric type (two types are not allowed in the same column).

5.5.2 Transcripts

TPM

Header: column 1 is "ID", column 2~N is "sample name", the sample name supports English letters, numbers and underscores, and does not support spaces and other special characters;

Column 1: transcript ID;

Columns 2~N are: TPM

FPKM

Header: column 1 is "ID", column 2~N is "sample name", the sample name supports English letters, numbers and underscores, and does not support spaces and other special characters;

Column 1: transcript ID;

Columns 2~N are: FPKM;

Read counts

Header: column 1 is "ID", column 2~N is "sample name", the sample name supports English letters, numbers and underscores, and does not support spaces and other special characters;

Column 1: transcript ID;

Columns 2~N are: read counts;

other

Header: column 1 is "ID", column 2~N is "sample name", the sample name supports English letters, numbers and underscores, and does not support spaces and other special characters;

Column 1: transcript ID;

Column 2~N: Support character type or numeric type (two types are not allowed in the same column).

5.5.3 Protein

Expression

Header: column 1 is "ID", column 2~N is "sample name", the sample name supports English letters, numbers and underscores, and does not support spaces and other special characters;

For protein project reports, column 1 is the protein ID (NCBI or Uniprot) used in this report, refer to the protein ID in the table on the home page of the report; For non-protein item reports or newly added items, column 1 is the NCBI protein ID.

Column 2 ~ N column: protein expression.

other

Header: column 1 is "ID", column 2~N is "sample name", the sample name supports English letters, numbers and underscores, and does not support spaces and other special characters;

For protein project reports, column 1 is the protein ID (NCBI or Uniprot) used in this report, refer to the protein ID in the table on the home page of the report; For non-protein item reports or newly added items, column 1 is the NCBI protein ID.

Column 2~N: Support character type or numeric type (mixed types are not allowed in the same column).