

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-dimensional 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

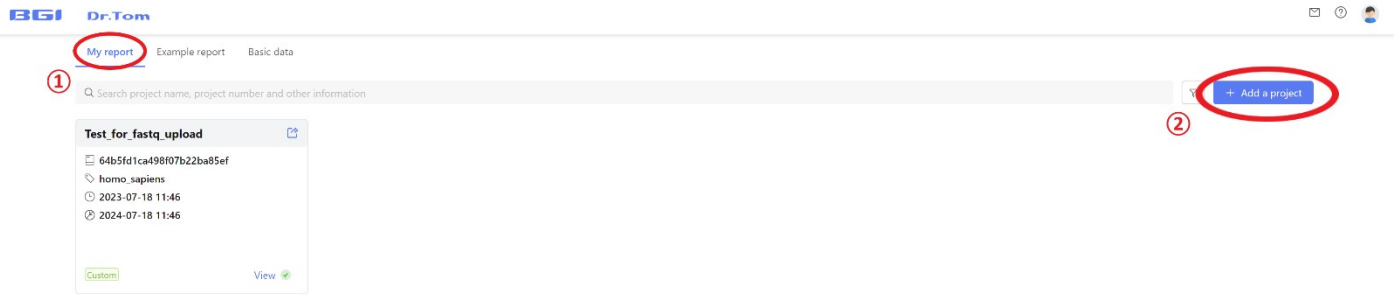
Data table can be prepared with Microsoft Excel.

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)

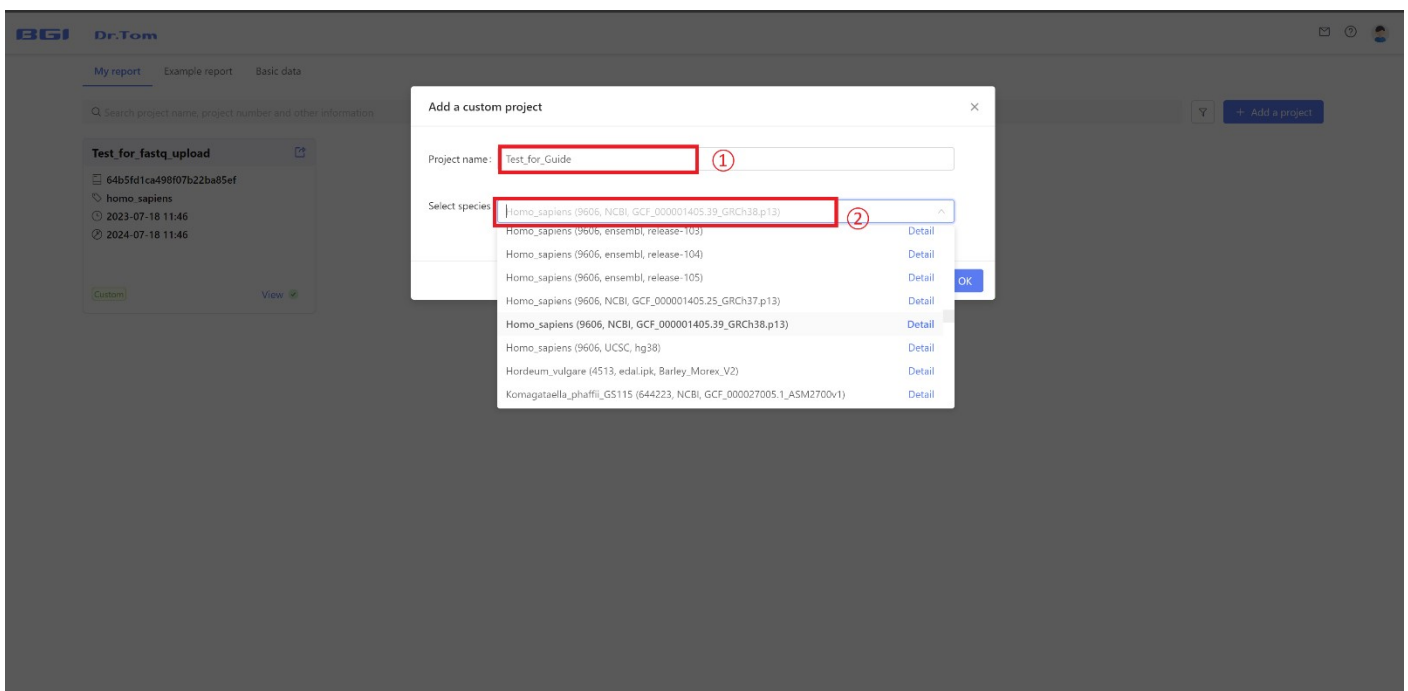
Data is suggested to be saved as a tab-delimited text file. Select "File -> Save As", for the file type, select "Text File (Tab-Delimited) (*.txt)".

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.

The screenshot shows the Dr.Tom interface with two project cards. The first card, 'Test_for_Guide', has a 'View' button circled in red. The second card, 'Test_for_fastq_upload', also has a 'View' button. A search bar at the top contains the text 'Search project name, project number and other information' and an 'Add a project' button.

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

The screenshot shows the Dr.Tom 'Upload' page. The 'upload' button is circled in red. Below the search bar is a table with columns for Gene ID, Gene Symbol, Type, KEGG Pathway Desc, KEGG Disease Desc, KEGG Module Desc, KEGG Reaction Desc, GO_C Desc, GO_F Desc, GO_P Desc, and CellMarker Desc. The table contains 15 rows of data.

Gene ID	Gene Symbol	Type	KEGG Pathway Desc	KEGG Disease Desc	KEGG Module Desc	KEGG Reaction Desc	GO_C Desc	GO_F Desc	GO_P Desc	CellMarker Desc
1	A1BG	mRNA	NA	NA	NA	NA	GO:0005576//e...	GO:0003674//...	GO:0002576//p...	A1BG
10	NAT2	mRNA	00232//Caffeine metabolism ...	NA	NA	R07940//No_INFO 1,7-Dimethylx...	GO:0005829//c...	GO:0004060//a...	GO:0006805//x...	NA
100	ADA	mRNA	00230//Purine metabolism 0...	H00092//T-B-Severe combined...	NA	R01560//Adenosine aminohydro...	GO:0005615//e...	GO:0001883//p...	GO:0000255//a...	ADA
1000	CDH2	mRNA	04514//Cell adhesion molecule...	H00293//Arrhythmogenic right...	NA	NA	GO:0005731//c...	GO:0005509//c...	GO:0003323//f...	CDH2
10000	AKT3	mRNA	01521//EGFR tyrosine kinase in...	H00027//Ovarian cancer H00...	NA	NA	GO:0005634//n...	GO:0000166//n...	GO:0000002//f...	AKT3
100008586	GAGE12F	mRNA	NA	NA	NA	NA	GO:0005575//c...	GO:0003674//...	GO:0008150//b...	NA
100009613	LINC02584	lncRNA	NA	NA	NA	NA	NA	NA	NA	NA
100009667	POU5F1P5	lncRNA	NA	NA	NA	NA	NA	NA	NA	NA
100009676	ZBTB11-AS1	lncRNA	NA	NA	NA	NA	NA	NA	NA	NA
10001	MED6	mRNA	NA	NA	NA	NA	GO:0000151//u...	GO:0003677//D...	GO:0006357//f...	MED6
10002	NR2E3	mRNA	NA	H00527//Retinitis pigmentosa...	NA	NA	GO:0005634//n...	GO:0000978//R...	GO:0000122//n...	NR2E3
10003	NAALAD2	mRNA	NA	NA	NA	NA	GO:0005886//p...	GO:0003824//c...	GO:0006508//p...	NAALAD2
100033407	VN2R19P	lncRNA	NA	NA	NA	NA	NA	NA	NA	NA

(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.

Dr.Tom

Test_for_Guide [Check history](#)

Selected species: Homo sapiens (9606, NCBI, GCF_000001405.39, GRCh38.p13)

Select upload content

Gene Transcript

> A set of related tools

Upload files [Check file description](#)

Please upload the tab separated text file, you can check the file in the “check history” after inputting successfully, or you can also use the tools. Please see the “check file description” for details.

NCBI Gene ID Read counts test.txt

upload

Fill in the plan [Completion Instructions](#)

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

Dr.Tom

Test_for_Guide [Check history](#)

Selected species: Homo sapiens (9606, NCBI, GCF_000001405.39, GRCh38.p13)

Select upload content

Gene Transcript

> A set of related tools

Upload files [Check file description](#)

Please upload the tab separated text file, you can check the file in the “check history” after inputting successfully, or you can also use the tools. Please see the “check file description” for details.

NCBI Gene ID Read counts

Fill in the plan [Completion Instructions](#)

The uploaded file does not meet the requirements for filling in the plan, please refer to the filling instructions on the right for details

Dr.Tom

Test for Guide

Check history

History of uploading records

Serial number	Upload number	Type	Number of valid records	Upload time	Entry status	Status	Detail
1	6500148ba498f06ca7e48cf4	Gene	100 row * 6 column	2023-09-12 06:34:35	Success	Show	View

Gene

Transcript

> A set of related tools

Upload files

Check file description

Please upload the tab separated text file, you can check the file in the "check history" after inputting successfully, or you can also use the tools. Please see the "check file description" for details.

NCBI Gene ID

Read counts

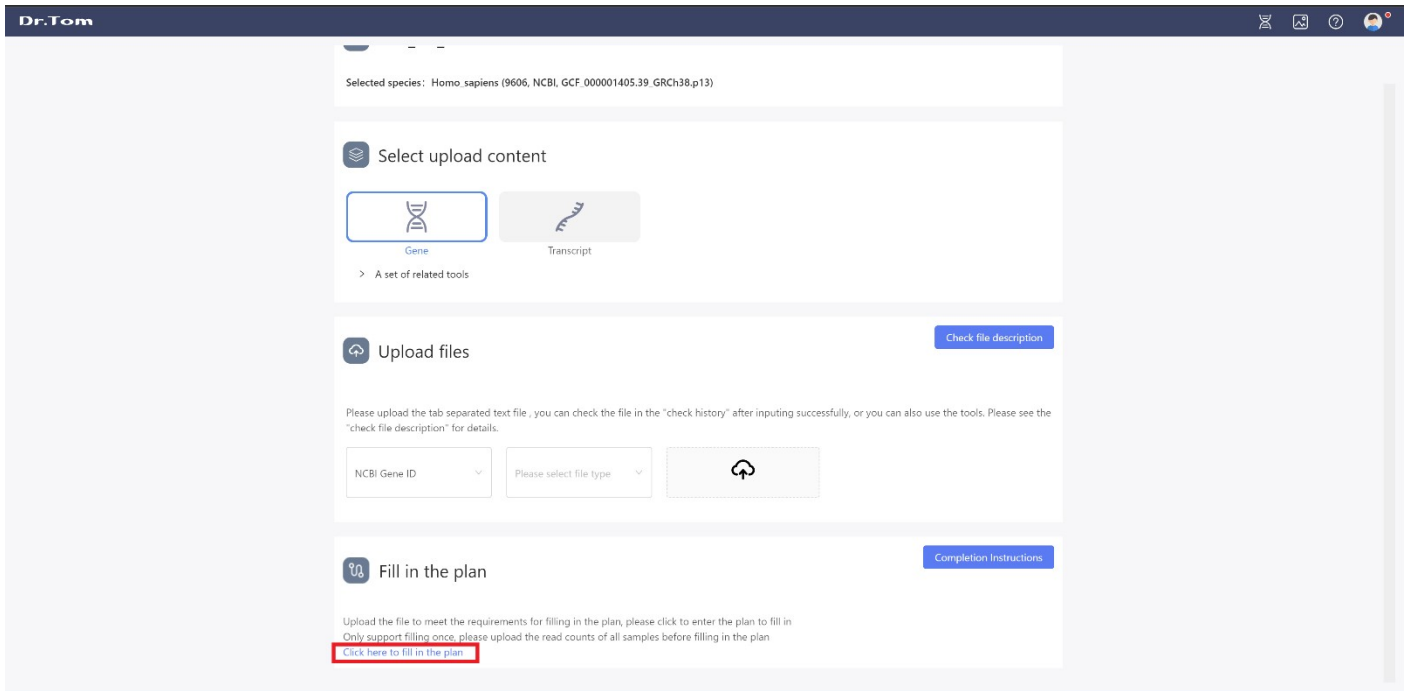
Fill in the plan

Completion Instructions

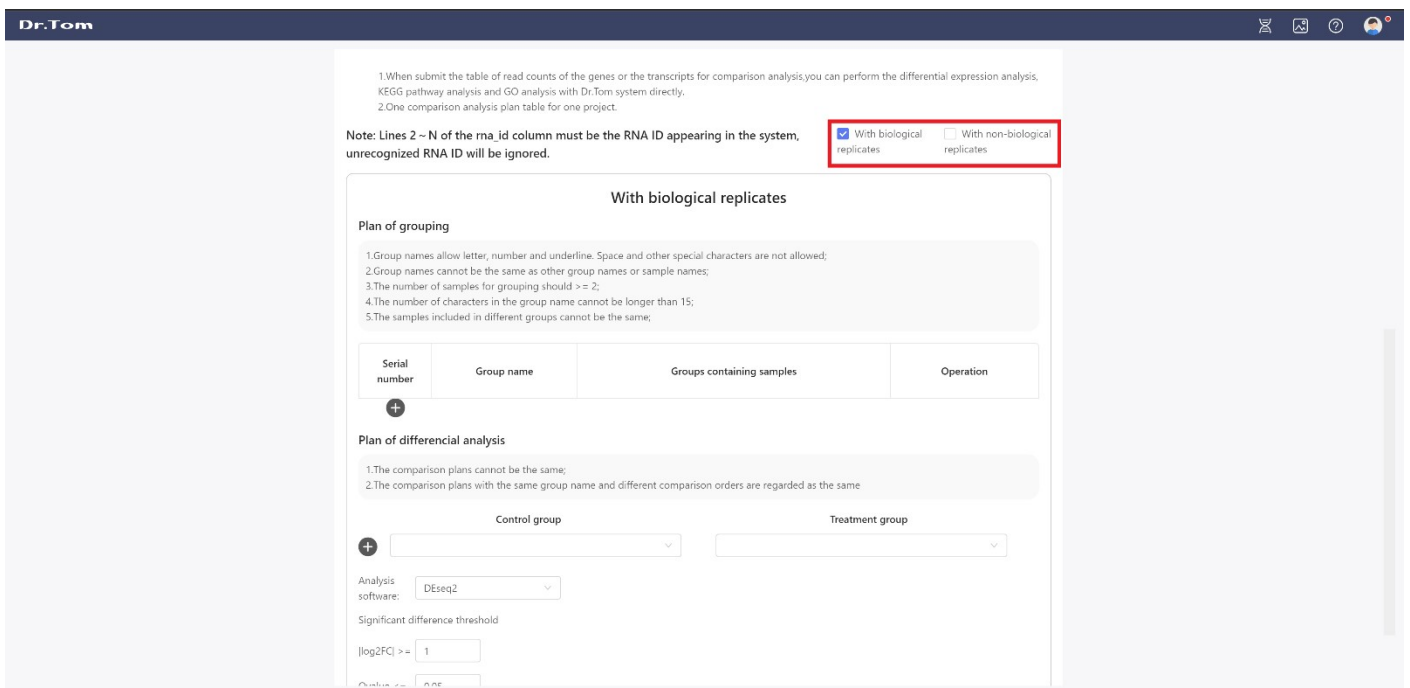
Upload the file to meet the requirements for filling in the plan, please click to enter the plan to fill in. Only support filling once, please upload the read counts of all samples before filling in the plan.

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.

1. When submit the table of read counts of the genes or the transcripts for comparison analysis, you can perform the differential expression analysis, KEGG pathway analysis and GO analysis with Dr.Tom system directly.
2. One comparison analysis plan table for one project.

Note: Lines 2 - N of the rna_id column must be the RNA ID appearing in the system, unrecognized RNA ID will be ignored. With biological replicates With non-biological replicates

With biological replicates

Plan of grouping

- Group names allow letter, number and underline. Space and other special characters are not allowed;
- Group names cannot be the same as other group names or sample names;
- The number of samples for grouping should ≥ 2 ;
- The number of characters in the group name cannot be longer than 15;
- The samples included in different groups cannot be the same;

Serial number	Group name	Groups containing samples	Operation
+			

Plan of differential analysis

- The comparison plans cannot be the same;
- The comparison plans with the same group name and different comparison orders are regarded as the same

Control group: Treatment group:

Analysis software:

Significant difference threshold:

1. When submit the table of read counts of the genes or the transcripts for comparison analysis, you can perform the differential expression analysis, KEGG pathway analysis and GO analysis with Dr.Tom system directly.
2. One comparison analysis plan table for one project.

Note: Lines 2 - N of the rna_id column must be the RNA ID appearing in the system, unrecognized RNA ID will be ignored. With biological replicates With non-biological replicates

With biological replicates

Plan of grouping

- Group names allow letter, number and underline. Space and other special characters are not allowed;
- Group names cannot be the same as other group names or sample names;
- The number of samples for grouping should ≥ 2 ;
- The number of characters in the group name cannot be longer than 15;
- The samples included in different groups cannot be the same;

Serial number	Group name	Groups containing samples	Operation
1	Group_A	<input type="text" value="Sample_1 Read Count"/> <input type="text" value="Sample_2 Read Count"/> <input type="text" value="Sample_3 Read Count"/> <input type="text" value="Sample_4 Read Count"/> <input type="text" value="Sample_5 Read Count"/> <input type="text" value="Sample_6 Read Count"/>	Delete
+			

Plan of differential analysis

- The comparison plans cannot be the same;
- The comparison plans with the same group name and different comparison orders are regarded as the same

Control group: Treatment group:

Analysis software:

Significant difference threshold:

1. When submit the table of read counts of the genes or the transcripts for comparison analysis, you can perform the differential expression analysis, KEGG pathway analysis and GO analysis with Dr.Tom system directly.
2. One comparison analysis plan table for one project.

Note: Lines 2 - N of the rna_id column must be the RNA ID appearing in the system, unrecognized RNA ID will be ignored. With biological replicates With non-biological replicates

With biological replicates

Plan of grouping

- Group names allow letter, number and underline. Space and other special characters are not allowed;
- Group names cannot be the same as other group names or sample names;
- The number of samples for grouping should ≥ 2 ;
- The number of characters in the group name cannot be longer than 15;
- The samples included in different groups cannot be the same;

Serial number	Group name	Groups containing samples	Operation
1	Group_A	<input type="text" value="Sample_1 Read Count"/> × <input type="text" value="Sample_2 Read Count"/> × <input type="text" value="Sample_3 Read Count"/> ×	Delete
2	Group_B	<input type="text" value="Sample_4 Read Count"/> × <input type="text" value="Sample_5 Read Count"/> ×	Delete
3	Group_C	<input type="text" value="Sample_6 Read Count"/> × <input type="text" value="Sample_5 Read Count"/> ×	Delete
+			

Plan of differential analysis

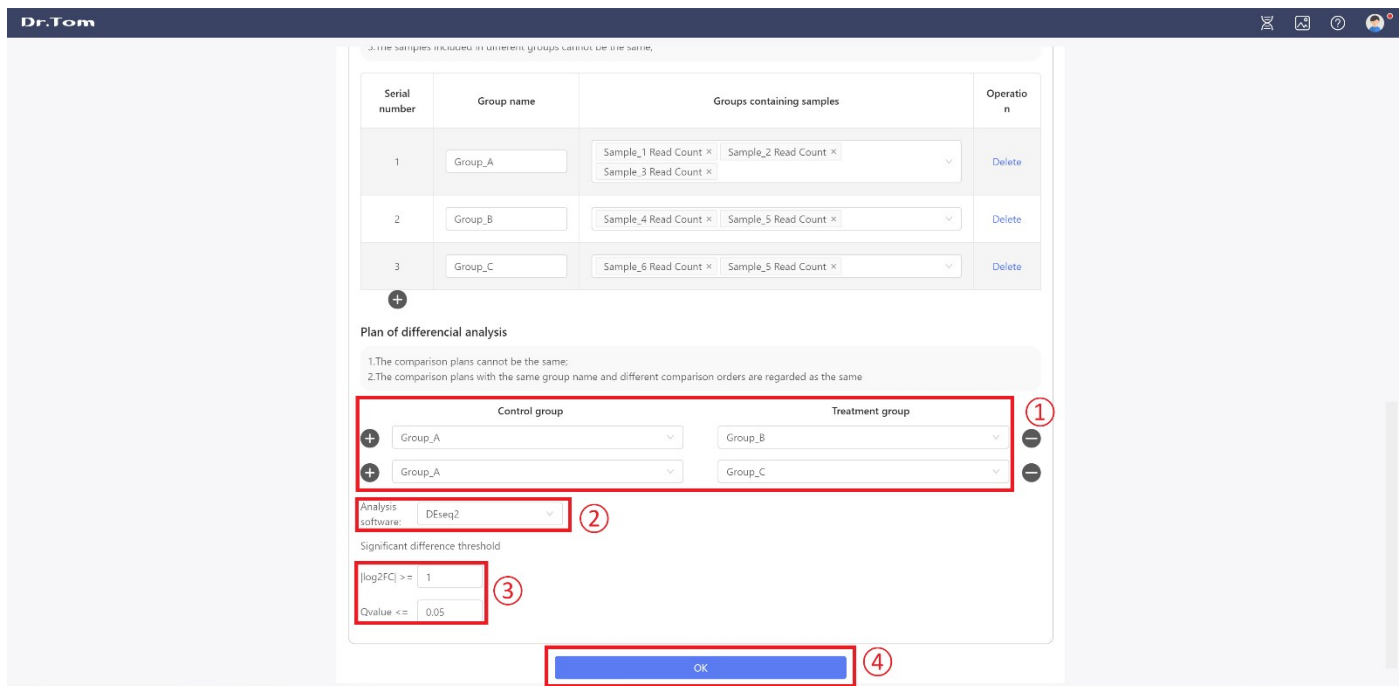
- The comparison plans cannot be the same;
- The comparison plans with the same group name and different comparison orders are regarded as the same

Control group: Treatment group:

(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: “ $|\log_2FC| \geq$ ” and “ $Q\text{-value} \leq$ ”; Click “OK” to submit analysis.

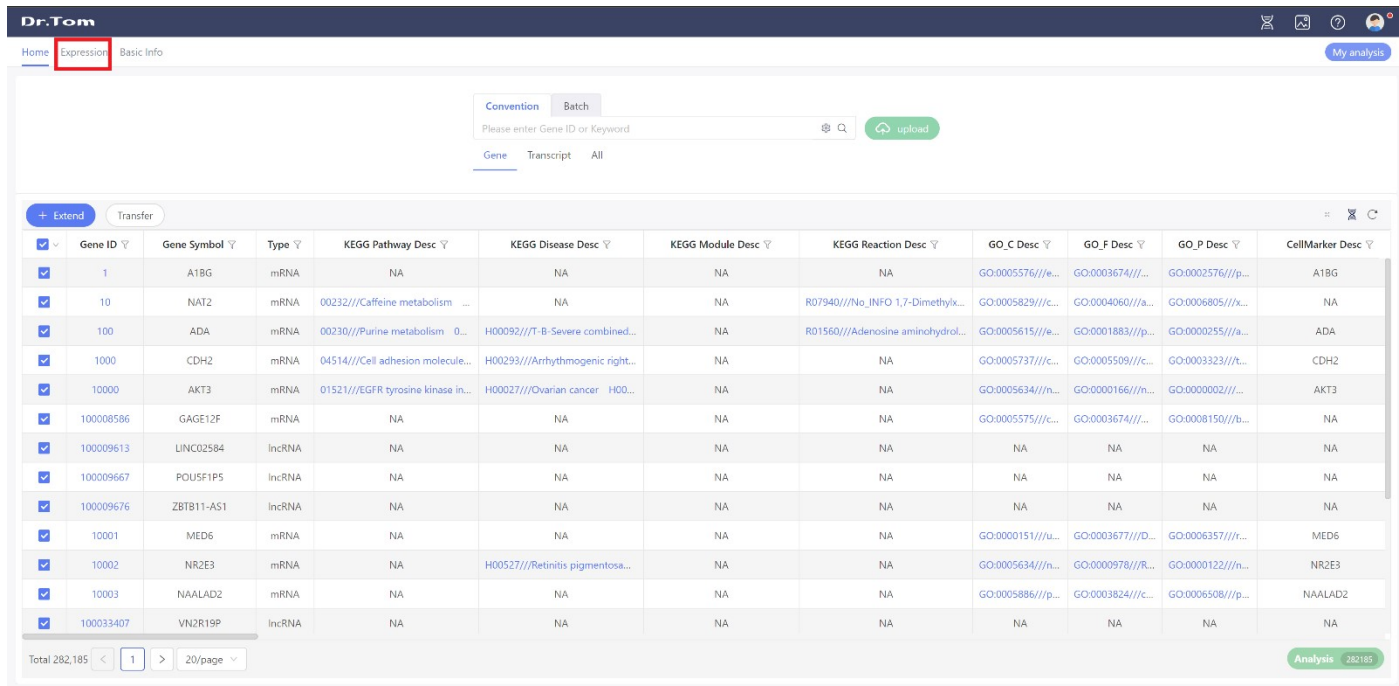
Notice: genes/transcripts with $|\log_2FC|$ lower than threshold and with Q-value higher than threshold will be filtered and not included in the result. Users is suggested to set the initial threshold with a moderate threshold or re-perform differential analysis later with analysis tool panel.



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).