Dr. Tom Data Upload

File Requirement

The system supports users to upload their own data for analysis. The upload contents include genes, transcripts, proteins, DMR or other information. If Differential Expression Gene Analysis (DEG analysis) is needed, **Read Count** must be uploaded while other types of data DO NOT support DEG analysis. The format of uploading files has specific requirements as follows:

1. Basic Requirements:

About file names

- File type: Please upload a Tab-delimited text file. After editing in Excel, select "File -> Save As -> File Type and select "Text File (Tab-delimited) (*.txt)" , Save;
- Sample name: It is recommended **not to exceed 15 characters**, otherwise it may cause problems such as occlusion of the sample legend when generating the graph.

File size

- It is required to be **less than 20M** for one file.

Steps of uploading your data

- Select the content of the data, the available options are gene, transcript, protein, DMR;
- After selecting the file type, follow the system prompts to complete the upload and submit;
- After the upload is successful, the system will check the data. After passing the checking, you can choose to view the data in the "Extension Column (in the table of homepage) -> User Upload" or perform analysis using analysis tools.

About ID

The upload ID needs to be consistent with the ID in the first column of the report (project) homepage table:

- When selecting "Gene", it mainly comes from NCBI (ID is generally a pure number), and it may also come from other public databases. For the specific source of ID, please refer to the information on the head of uploading page. Some species supports Gene Symbol and Ensembl Gene ID upload and you can check it when selecting ID type;

- When selecting "transcript", it mainly comes from NCBI (ID generally starts with "NM_" and "XM_"), or it may come from other public databases. For the specific source of ID, please refer to the information on the head of uploading page.
- When selecting "protein", it mainly comes from NCBI or Unitprot. The ID should be from one database if there are several files need to be uploaded in one project.
- When "DMR" is selected, the DMR ID named by this system and custom DMR ID are supported (location information needs to be uploaded).

1. File format

A. Gene

TPM

- Header: First column should be "ID". The 2nd and subsequent column should be "Sample name", such as “sample1”. The sample name supports English letters, numbers and underscores, and does not support spaces and other special characters;
- The content of the first column is: Gene ID;
- Columns 2~N are: gene expression TPM value;

Example:

ID	sample1	sample2	sample3	sample4	sample5	... ...
54904	41.31	77.42	81.31	65.72	44.07	
3575	12	19.03	16.78	25.67	11.79	

FPKM

- Header: First column should be "ID". The 2nd and subsequent column should be "Sample name", such as “sample1”. The sample name supports English letters, numbers and underscores, and does not support spaces and other special characters;
- The content of the first column is: Gene ID;
- Columns 2~N are: gene expression FPKM value;

Example:

ID	sample1	sample2	sample3	sample4	sample5	... ...
54904	9.07	4.44	7.79	19.04	20.39	
3575	23.5	16.25	20.54	18.47	21.64	

Read counts

- Header: First column should be "ID". The 2nd and subsequent column should be "Sample name", such as “sample1”. The sample name supports English letters, numbers and underscores, and does not support spaces and other special characters;
- The content of the first column is: Gene ID;
- Columns 2~N are: gene expression Read count value;

Example:

ID	sample1	sample2	sample3	sample4	sample5	... ...
54904	4986.54	4310.81	5324.52	3418.99	3599.76	
3575	568.97	426.88	449.29	222.13	313.36	

Others

- Header: First column should be "ID". The 2nd and subsequent column should be "Sample name", such as “sample1”. The sample name supports English letters, numbers and underscores, and does not support spaces and other special characters;
- The content of the first column is: Gene ID;
- Columns 2~N supports character type or number type data but only one type of data can be allowed in one column.

Example:

ID	test1	test2	test3	test4	test5	... ...
2817	12.05	71.24	22.67	45.19	34.38	
9817	13.88	60.45	20.39	25.21	64.97	

2. Transcript

TPM

- Header: First column should be "ID". The 2nd and subsequent column should be "Sample name", such as “sample1”. The sample name supports English letters, numbers and underscores, and does not support spaces and other special characters;
- The content of the first column is: Transcript ID;
- Columns 2~N are: transcript expression TPM value;

Example:

ID	sample1	sample2	sample3	sample4	sample5	... ...
NM_006851.2	41.31	77.42	81.31	65.72	44.07	
NM_005190.3	12	19.03	16.78	25.67	11.79	

FPKM

- Header: First column should be "ID". The 2nd and subsequent column should be "Sample name", such as “sample1”. The sample name supports English letters, numbers and underscores, and does not support spaces and other special characters;
- The content of the first column is: Transcript ID;
- Columns 2~N are: transcript expression FPKM value;

Example:

ID	sample1	sample2	sample3	sample4	sample5	... ...
NM_006851.2	9.07	4.44	7.79	19.04	20.39	
NM_005190.3	23.5	16.25	20.54	18.47	21.64	

Read counts

- Header: First column should be "ID". The 2nd and subsequent column should be "Sample name", such as “sample1”. The sample name supports English letters, numbers and underscores, and does not support spaces and other special characters;
- The content of the first column is: Transcript ID;
- Columns 2~N are: transcript expression Read Count value;

Example:

ID	sample1	sample2	sample3	sample4	sample5	... ...
NM_006851.2	4986.54	4310.81	5324.52	3418.99	3599.76	
NM_005190.3	568.97	426.88	449.29	222.13	313.36	

Others

- Header: First column should be "ID". The 2nd and subsequent column should be "Sample name", such as “sample1”. The sample name supports English letters, numbers and underscores, and does not support spaces and other special characters;
- The content of the first column is: Gene ID;
- Columns 2~N supports character type or number type data but only one type of data can be allowed in one column.

Example:

ID	test1	test2	test3	test4	test5	... ...
NM_006851.2	12.05	71.24	22.67	45.19	34.38	
NM_005190.3	13.88	60.45	20.39	25.21	64.97	

3. Protein

Expression

- Header: First column should be "ID". The 2nd and subsequent column should be "Sample name", such as “sample1”. The sample name supports English letters, numbers and underscores, and does not support spaces and other special characters;
- The first column: protein ID.
  If users upload the protein data to an existed protein project, the version of protein ID should be consistent with the existed protein data; If users upload the protein data to a non-protein project or a new project, the protein ID comes NCBI.
- Columns 2~N are the expression value of protein.

Example:

ID	sample1	sample2	sample3	sample4	sample5	... ...
sp|Q9Z2P6|SNP29_RAT	21.09	18.88	26.11	10.03	16.87	
sp|Q9Z2M4|DECR2_RAT	19.23	31.92	15.33	12.89	14.66	

Others

- Header: First column should be "ID". The 2nd and subsequent column should be "Sample name", such as “sample1”. The sample name supports English letters, numbers and underscores, and does not support spaces and other special characters;
- The content of the first column is: Protein ID;
- Columns 2~N supports character type or number type data but only one type of data can be allowed in one column.

Example:

ID	name1	name2	name3	name4	name5	... ...
sp|Q9Y6W5|WASF2_HUMAN	12.05	71.24	22.67	45.19	34.38	
sp|Q9Y6E2|BZW2_HUMAN	13.88	60.45	20.39	25.21	64.97	

4. DMR

DMR upload supports the existing DMR ID of this project and the newly entered customized DMR ID. The customized ID needs to upload location information according to the format requirements.

Add customized DMR

- Column 1: dmr_id, which is named by the non-BGI system for the first upload;
- Column 2: chr, which indicates the chromosome number;
- Column 3: start, which represents the starting position;
- Column 4: end, which indicates the end position;
- Column 5: context, which indicates the type of methylation, such as CG, CHG, CHH;
- Column 6 ~ Column N: extended information, such as annotation information (characters and numbers are not allowed in the same column);

Example:

dmr_id	chr	start	end	context	... ...
DMR001	chr10	1000001	1000201	CG	
DMR002	chr11	1000389	1000589	CHG	

Supplement DMR information

- Column 1: dmr_id, which is the ID that has been uploaded in this project;
- Column 2~Column N: Support character type or number type (two types are not allowed in the same column);
- Sample names, group names support English letters, numbers and underscores, spaces and other special characters are not supported

Example:

dmr_id	Anno Info
chr2_192576801_192577000	Intergenic:chr2:192197934-195572943
chr2_192570601_192570800	MLT1F2:chr2:192570709-192571291Intergenic:chr2:192197934-195572943