Dr. Tom-文件上传要求

本系统支持用户上传自有数据进行分析，上传内容包括基因、转录本、蛋白、DMR 信息。**若需进行差异分析，需上传 Read Counts 文件，其他文件类型不支持进行差异分析。**上传的文件需符合一定要求，具体如下:

一、基本要求:

关于名称

- 文件扩展名：**请上传制表符分隔的文本文件**，可以在 Excel 中完成编辑后，依次选择“文件 -> 另存为 -> 文件类型选择“文本文件（制表符分隔）（*.txt）”，完成保存;
- 样品名：**建议不超过 15 个字符**，否则生成结果图片时可能造成样品图例遮挡等问题。

文件大小

- 要求单次上传文件需**小于 20M**。

上传步骤

- 选择需要上传的内容，可选项为基因、转录本、蛋白、DMR;
- 选择下方文件类型后，按照系统提示完成上传，并提交;
- 上传成功后，系统将对数据做校验，录入到数据中。录入成功后，可以在报告的“扩展列 -> 用户上传”和相应分析工具中选择上传数据进行查看或分析。

关于 ID

上传 ID 需要与系统使用的参考基因组版本一致：

- 选择“基因”时，主要来源于 NCBI（ID 一般为纯数字），也可能来自于其它公共数据库，ID 具体来源可查看数据上传页面上方的“已选择物种”信息。部分物种支持 Gene Symbol 和 Ensembl Gene ID 上传，可通过上传文件中的“ID 类型”选项，查看当前物种是否支持;
- 选择“转录本”时，主要来源于 NCBI（ID 一般为“NM_”、“XM_”开头），也可能来自于其它公共数据库，ID 具体来源可查看数据上传页面上方的“已选择物种”信息;
- 选择“蛋白”时，： ID 主要来自于 Uniprot 或 NCBI。若在同一项目中分批上传数据，上传蛋白 ID 需保持一致;
- 选择“DMR”时，支持本系统命名的 DMR ID，以及自定义 DMR ID（需上传坐标信息）

二、上传文件格式：

1. 基因

TPM

- 表头：第 1 列为“ID”，第 2~第 N 列为“样本名 ”，样本名支持英文字母、数字与下划线，不支持空格与其它特殊字符；
- 第 1 列为：基因 ID；
- 第 2~第 N 列为：基因表达量 TPM 值；

示例文件：

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

- 表头：第 1 列为“ID”，第 2~第 N 列为“样本名 ”，样本名支持英文字母、数字与下划线，不支持空格与其它特殊字符；
- 第 1 列为：基因 ID；
- 第 2~第 N 列为：基因表达量 FPKM 值；

示例文件：

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

- 表头：第 1 列为“ID”，第 2~第 N 列为“样本名 ”，样本名支持英文字母、数字与下划线，不支持空格与其它特殊字符；
- 第 1 列为：基因 ID；
- 第 2~第 N 列为：基因表达量 read counts 值；

示例文件：

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	

其它

- 表头：第 1 列为“ID”，第 2~第 N 列为“样本名 ”，样本名支持英文字母、数字与下划线，不支持空格与其它特殊字符；
- 第 1 列为：基因 ID；
- 第 2~第 N 列为：支持字符类型或数字类型（同一列中不允许同时出现两种类型）。

示例文件：

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. 转录本

TPM

- 表头：第 1 列为“ID”，第 2~第 N 列为“样本名 ”，样本名支持英文字母、数字与下划线，不支持空格与其它特殊字符；
- 第 1 列为：转录本 ID；
- 第 2~第 N 列为：基因表达量 TPM 值；

示例文件：

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

- 表头：第 1 列为“ID”，第 2~第 N 列为“样本名 ”，样本名支持英文字母、数字与下划线，不支持空格与其它特殊字符；
- 第 1 列为：转录本 ID；
- 第 2~第 N 列为：基因表达量 FPKM 值；

示例文件：

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

- 表头：第 1 列为“ID”，第 2~第 N 列为“样本名 ”，样本名支持英文字母、数字与下划线，不支持空格与其它特殊字符；
- 第 1 列为：转录本 ID；
- 第 2~第 N 列为：基因表达量 read counts 值；

示例文件：

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	

其它

- 表头：第 1 列为“ID”，第 2~第 N 列为“样本名 ”，样本名支持英文字母、数字与下划线，不支持空格与其它特殊字符；
- 第 1 列为：转录本 ID；
- 第 2~第 N 列为：支持字符类型或数字类型（同一列中不允许同时出现两种类型）。

示例文件：

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. 蛋白

Expression

- 表头：第 1 列为“ID”，第 2~第 N 列为“样本名 ”，样本名支持英文字母、数字与下划线，不支持空格与其它特殊字符；
- 若为蛋白项目报告，第 1 列为**本报告中使用的蛋白 ID（NCBI 或 Uniprot）**，参考报告主页表格中的蛋白 ID；若为非蛋白项目报告或新添加项目，**第 1 列为 NCBI 蛋白 ID**。
- 第 2 列~第 N 列为：蛋白表达量 XXX 值。
-

示例文件：

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	

其它

- 表头：第 1 列为“ID”，第 2~第 N 列为“样本名 ”，样本名支持英文字母、数字与下划线，不支持空格与其它特殊字符；
- 若为蛋白项目报告，第 1 列为**本报告中使用的蛋白 ID（NCBI 或 Uniprot）**，参考报告主页表格中的蛋白 ID；若为非蛋白项目报告或新添加项目，**第 1 列为 NCBI 蛋白 ID**。
- 第 2~第 N 列为：支持字符类型或数字类型（同一列中不允许同时出现两种类型）。

示例文件:

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 上传支持本项目已有的 DMR ID 以及新录入的自定义 DMR ID，自定义 ID 需按照格式的要求上传坐标信息。

新增自定义 DMR

- 第 1 列：dmr_id，此处 ID 为首次上传非华大系统命名的 ID;
- 第 2 列：chr 表示位于染色体编号;
- 第 3 列：start 表示起始位置;
- 第 4 列：end 表示终止位置;
- 第 5 列：context 表示甲基化类型，如 CG、CHG、CHH;
- 第 6 列~第 N 列：扩展信息，例如注释信息（同一列中不允许同时出现字符与数字两种类型）;

示例文件:

dmr_id	chr	start	end	context	... ...
DMR001	chr1	1000201	1000401	CG	
DMR002	Chr4	1000780	1000980	CHH	

补充 DMR 属性

- 表头：第 1 列为“dmr_id”，第 2~第 N 列为“属性名称 ”，样本名支持英文字母、数字与下划线，不支持空格与其它特殊字符；
- 第 1 列：dmr_id，此处 ID 为已在本项目中上传过的 ID
- 第 2 列~第 N 列：支持字符类型或数字类型（同一列中不允许同时出现两种类型）；
- 样品名、组名支持英文字母、数字与下划线，不支持空格与其它特殊字符。

示例文件：

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