How to create a target file for Shaman

format, contents and constraints

The target file is required for each analysis done with Shaman. It must contain all the available information on the samples (corresponding to the metadata) that will be used to build the statistical model and/or to visualize the data. To be loaded in Shaman the target file must respect some properties

1. **The first column is dedicated to the sample name which must correspond exactly to the column names of the count matrix. At least 2 samples are required.** *Once the count matrix is loaded, check carefully the sample names in the “count table” tab. Sometimes some characters are modified with the loading.*
2. **At least one variable must be provided.** *In Example 1, two variables are provided (condition and treatment).*
3. **NA or missing values are not allowed.**
4. **A variable with the same value for each sample is not allowed**. *This kind of variable should be removed from the target file before loading.*
5. **The selected variables for the statistical model must not be collinear**. *It means that if one variable can be determined by another variable or a combination of variables the analysis cannot be done with all the variables. However, the user can user will be able to use this variable for visualization. (See example 3).*
6. **Be careful, numeric variables will be considered as quantitative variable.** *For instance, do not use 1 and 2 to describe two different conditions but C1 and C2 or A and B. (see exemple 3)*
7. **Avoid using special characters such as** _/\?*:<>|+,[]-+()’%@"&_

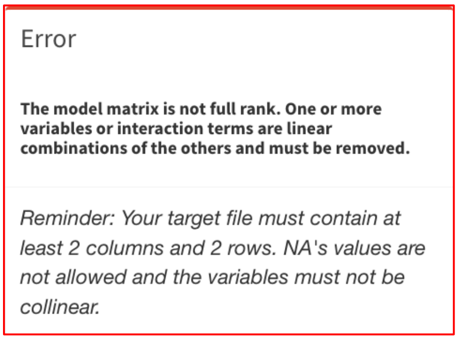
Example 1: Target file with 2 variables (condition and treatment)

sampleID	condition	treatment
S1	WT	A
S2	WT	A
S3	KO	A
S4	KO	A
S5	WT	B
S6	WT	B
S7	KO	B
S8	KO	B

This is a usual example in which we have 2 variables to describe the samples (condition and treatment). For instance, the user will be able to define the following model: condition + treatment + condition:treatment and then get differentially abundant features between treatments A and B for each condition.

Example 2: Target file with collinearity problem

sampleID	condition	treatment	group
S1	WT	A	g1
S2	WT	A	g1
S3	KO	A	g2
S4	KO	A	g2
S5	WT	B	g3
S6	WT	B	g3
S7	KO	B	g4
S8	KO	B	g4


Error

The model matrix is not full rank. One or more variables or interaction terms are linear combinations of the others and must be removed.

Reminder: Your target file must contain at least 2 columns and 2 rows. NA's values are not allowed and the variables must not be collinear.

In this example, group = condition + treatment, so the variables are collinear.

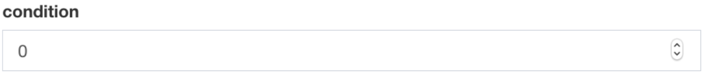
Note that this file can be loaded in Shaman without error but the error will appear if the user tries to define a model with the three variables condition, treatment and group.

Example 3: Quantitative versus qualitative variable.

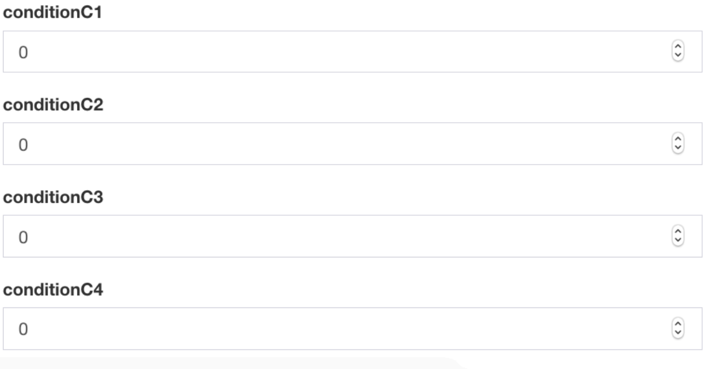
Target file

Model parameters

sampleID	condition
S1	1
S2	1
S3	2
S4	2
S5	3
S6	3
S7	4
S8	4


condition: 0

sampleID	condition
S1	C1
S2	C1
S3	C2
S4	C2
S5	C3
S6	C3
S7	C4
S8	C4


conditionC1: 0

conditionC2: 0

conditionC3: 0

conditionC4: 0

In case 1, condition is considered as a numeric variable which leads to only one parameter in the statistical model. It assumes that the difference between 1 and 3 is two times the difference between 1 and 2 and so on. In case 2, there is no order between the conditions.