Preparing your data

Researchers are responsible for providing the data to be analysed in a suitable format. Below you can find instructions to prepare the data. If you have any questions, just contact us via biostatistics@nki.nl.

General instructions

Anonymised data: All data must be anonymised, with sample ids representing labels that can only be connected to individuals via a table that is not provided to us.

Tab-delimited data files: The data should reach us in a tab- or comma-delimited format.

Decimal separator: Use as decimal separator the dot '.', not comma. Specifically, when exporting data from excel in a computer that uses the Dutch standard, often the comma is used as decimal separator. Either set the standard used for the decimal separator as the dot, or open the file in a text editor (such as Notepad in Windows) and replace all commas by dots.

Missing values: missing values should preferably be indicated via an 'NA' in the corresponding entry, or else left empty (so no space).

Mouse and clinical studies data

Datasets from mouse/clinical studies should be organized in a table form, with one row per individual and one column per variable. If categorical variables are included, classes corresponding to codes used must be provided in a separate file. There should be one file per table or dataset.

Omics data

For studies involving omics data, *per omics data type* the following (tab- or comma-delimited) files should be provided: the omics data table with features on rows and samples on columns, with unique row identifiers (ideally in the first column) as well as unique column identifiers; the annotation data with features on rows and annotation variables on columns, with the same type of row identifiers as the omics data, and ideally the same number of rows as the omics data; and a phenotypic data table, which is a file with samples on rows and sample annotation variables on columns, including one column with the column identifiers of the omics data file. If multiple types of omics data are provided, then a single phenotypic data table may be used, where the column identifiers of each omics data file is included as a separate variable.

For studies using genetic screens (for example CRISPR screens), follow similar instructions as those for omics data, except that the omics data table is replaced by the screen data.

Examples:

1. Clinical data

Sample	Variable1	Factor1	Response1
s1	0.35	1	57
s2	0.43	2	101
s3	0.97	NA	76

Then in a separate text file:

Variable “Factor1” includes categories:
1 = age 15-29
2 = age 30-49

2. Omics data

Omics data table

ID	myfilex1.bam	myfilex2.bam	myfilex3.bam
gene1	534	10001	967
gene2	399	673	371
gene3	493	5992	591

Annotation data

ID	Chromosome	strand	start	end
gene1	1	+	13567000	13569183
gene2	19	-	2122337	NA
gene3	7	-	34781923	34782174

Phenotypic data table

Sample	Genotype	treatment
myfilex1.bam	WT	untreated
myfilex2.bam	WT	treated
myfilex3.bam	BRCA-	untreated