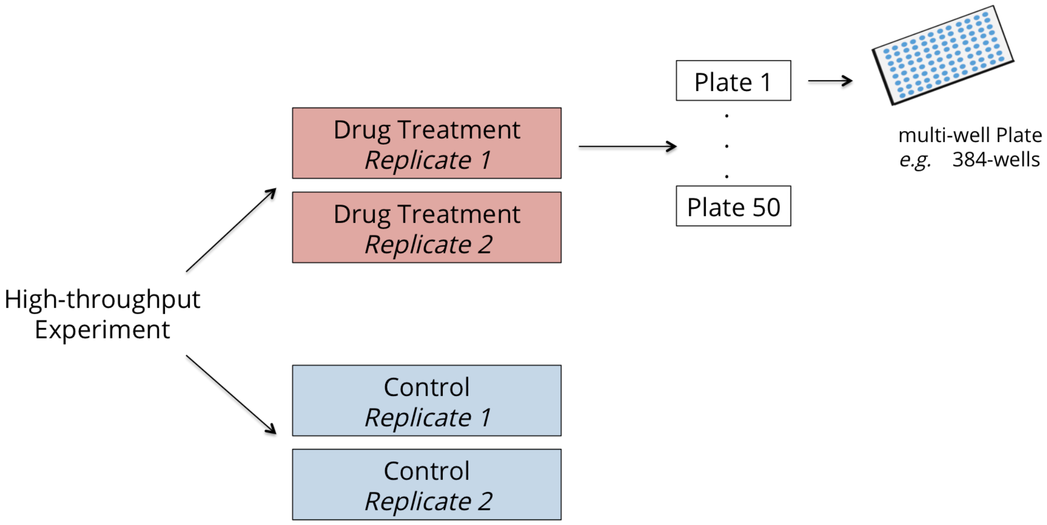
Supplement - Data format requirements

Global structure of the data set

HTSvis is a web application for the visualization of data from arrayed high-throughput screening (HTS) experiments. Such experiments typically spread over sets of plates, which are screened in replicates and possibly under different conditions (see Supplementary Figure 1). Formats supported by the application are 6-,12-,48-,96- and 384-well plates. Example data sets can be downloaded from the GitHub repository or via an external link.

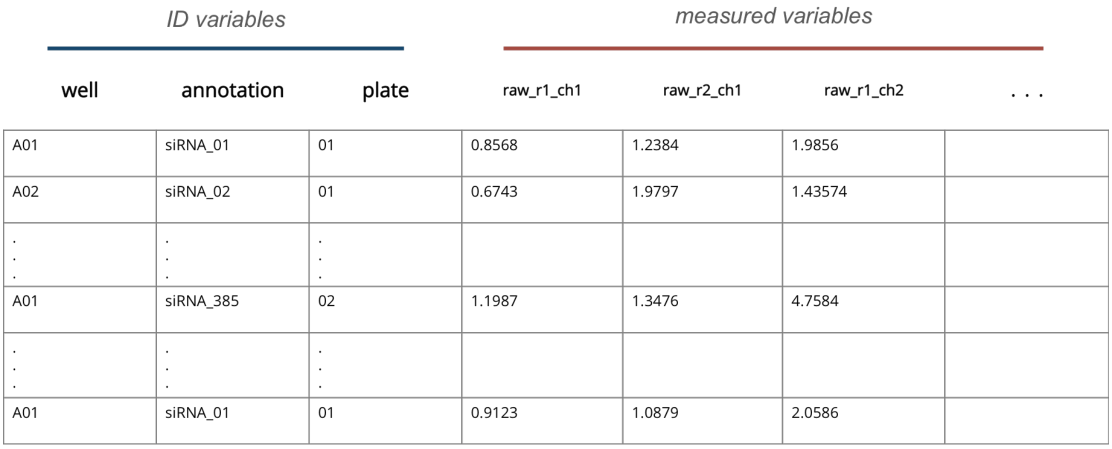


Supplementary Figure1: Schematic illustration of a large-scale experiment using multi-titer plates.

Data input requires a tabular format (.csv, .txt, .xlsx or .RData). The input data table can be either a result file from a statistical analysis using the Bioconductor/R package cellHTS or a generic spread sheet table (e.g. raw data). The two possible types of input tables, require different structures:

1. cellHTS result file (topTable.txt)

The cellHTS package provides a summary table which is exported as delimited text file (topTable.txt) at the final stage of analysis. This text file has a defined structure, fixed column names and can be loaded directly into HTSvis. Both, the structure and the column names of the *topTable* object should not be changed as HTSvis relies on those. The expected structure of the *topTable* is illustrated in Supplementary Figure 2.


ID variables
measured variables

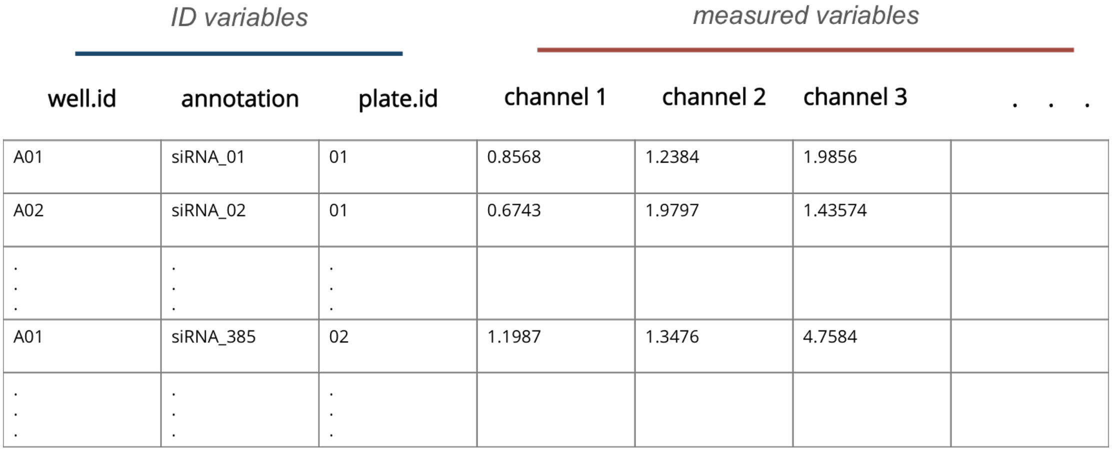
well	annotation	plate	raw_r1_ch1	raw_r2_ch1	raw_r1_ch2	. . .
A01	siRNA_01	01	0.8568	1.2384	1.9856	
A02	siRNA_02	01	0.6743	1.9797	1.43574	
. . .	. . .	. . .				
A01	siRNA_385	02	1.1987	1.3476	4.7584	
. . .	. . .	. . .				
A01	siRNA_01	01	0.9123	1.0879	2.0586	

Supplementary Figure 2: Schematic illustration of the cellHTS result file (topTable.txt).

The *well* and *plate* annotation columns are strictly required as also descried in the documentation of the cellHTS package (https://bioconductor.org/packages/release/bioc/html/cellHTS2.html). The annotation column is optional and can contain any kind of per-well annotation (e.g. gene symbols or reagent ids). The topTable will contain additional columns with data points assigned to each well. Depending on the design of the experiment, those columns contain measured values of replicates. Data from single and dual channel experiments can be loaded in the application. Besides raw measured values, further columns with normalized values and additional metrics are available.

2. Generic data table from arrayed screens

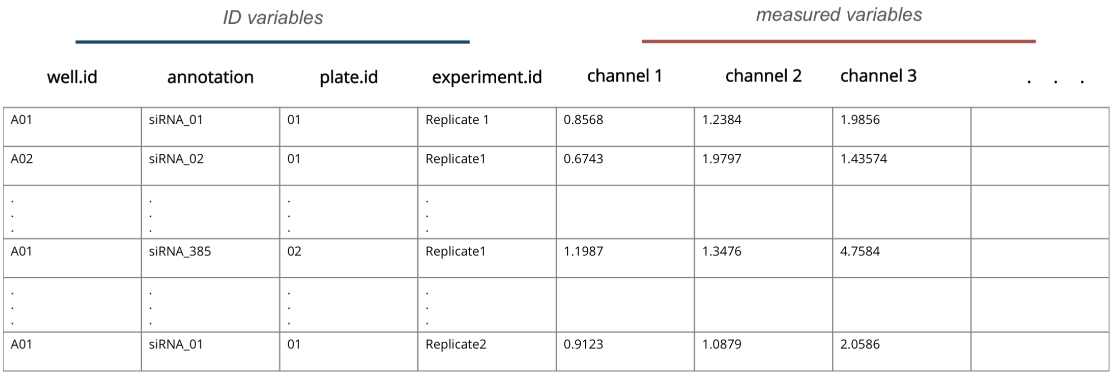
At least two identifier columns are required for a table to be loaded into HTSvis: the *well* and *plate* annotation. Those columns assign the sample values (termed *channels* in HTSvis) to the well position and the plate. Sample values per well are represented as column entries in the data table with one column per channel (Supplementary Figure 3). The number of channels per well is not limited. Columns containing additional information such as an annotation can be present. Accordingly, the data table contains row-wise entries per well as illustrated in Supplementary Figure 3. If an experiment contains multiple plates, the rows with *well* entries of each *plate* are pasted below each other. Since the user chooses the annotation columns with the corresponding spatial allocations from the user interface, column names can be named at will.



ID variables			*measured variables*			
well.id	annotation	plate.id	channel 1	channel 2	channel 3	. . .
A01	siRNA_01	01	0.8568	1.2384	1.9856	
A02	siRNA_02	01	0.6743	1.9797	1.43574	
. . .	. . .	. . .				
A01	siRNA_385	02	1.1987	1.3476	4.7584	
. . .	. . .	. . .				

Supplementary Figure 3: Schematic illustration of a generic table with the minimal required annotation to be read in HTSvis.

If the table contains individual sets of plates (e.g. replicates), each well has to be assigned to those. Accordingly, an additional annotation column has to be present. The required structure is illustrated below in Supplementary Figure 4. Importantly, in the case of such a data structure the individual sets have to contain the same collection of plates (same plate identifiers are required in each set)



ID variables				*measured variables*			
well.id	annotation	plate.id	experiment.id	channel 1	channel 2	channel 3	. . .
A01	siRNA_01	01	Replicate 1	0.8568	1.2384	1.9856	
A02	siRNA_02	01	Replicate1	0.6743	1.9797	1.43574	
. . .	. . .	. . .	. . .				
A01	siRNA_385	02	Replicate1	1.1987	1.3476	4.7584	
. . .	. . .	. . .	. . .				
A01	siRNA_01	01	Replicate2	0.9123	1.0879	2.0586	

Supplementary Figure 4: Schematic illustration of the structure required to read in a table with with multiple experiments. An additional column with the experiment allocation must be present.

Well- and file format

An important requirement concerning the data structure is that the dataset is symmetrical in a way that each experiment contains the same set of plates and each plate has the same well and complete format. This is especially relevant if multiple experiments are investigated because the plate identifiers need to be consistent between the experiments. Furthermore the data set has to be complete which means that all experiments need to contain the same number of plates. All plates further have to be complete in respect to the well format with numeric entries for all wells. Missing values (or such flagged during statistical analysis) should be filled with NA or NaN.

The well annotation has to follow an alphanumerical encoding with column positions indicated by numbers and row positions by letters. The specific row and column annotation format is: 'RowColmn' with rows as letters and columns as numbers (e.g. A1 or A01). Letters can be upper- or lowercase. Letters and numbers may be separated by characters like '-' or '_'. Delimited tables (.txt, .csv) should be uniformly separated by tab, comma, semicolon or space, tables saved as RData files must be data frames. All columns should be named.