GDSC raw data

GDSC raw data is available as a download from http://www.cancerrxgene.org/downloads. The dose response curve fitting uses the non-linear mixed effects model described in Vis, D.J. et al. Pharmacogenomics 2016, 17(7):691-700 (https://www.ncbi.nlm.nih.gov/pubmed/27180993). To fit the data with the model use the R package gdscIC50 (https://github.com/CancerRxGene/gdscIC50).

Experimental design

- GDSC drug screening uses 1536 well plates and 384 well plates.
- Each plate is plated with a single cell line.
- The plate layout, i.e., what happens in a given well position - is it a control, is it a drug treatment? - is called a drug set and is given a unique `DRUGSET_ID`.
- The screening plate tags within the drugset describe the treatment for a particular well. They allow wells to be grouped together by treatment type, e.g., drugged wells, control wells.
- The tags are used as part of the lab procedure to instruct the liquid handling robotics on how to array drug treatments in a particular plate (the drugset); and as part of downstream analysis (QC, dose-response fitting, etc.)
- There can be more than one tag per plate position. Thus in raw data files (csv) there may be more than one row per plate well position, e.g., `L12-D1-S + DMSO`.
- Single drug treatments are referred to as library drugs - `L1`,`L2` etc. Each library is used in a titration to elicit the dose response characteristics for that compound with a particular cell line.
- Some GDSC data includes combination drug treatments. These are usually a titration of a library compound in combination with a fixed dose anchor compound - `L1 + A1` etc. Combination treatment data are not currently available from www.cancerrxgene.org

For more details on the tags see below.

GDSC raw data format

GDSC raw data is distributed as a csv file which can then be loaded as a data frame. The `gdsc_example` dataset contains the minimum columns for the GDSC raw data to work wth the gdscIC50 package. Other GDSC data sets may contain additional columns. Not all well positions per plate are represented in public data sets because some drug treatments are part of private collaborations.

```
library(gdscIC50)
data("gdsc_example")
gdsc_example[99:100,]
    RESEARCH_PROJECT BARCODE SCAN_ID        DATE_CREATED
99            GDSC_SA    3230    2945 2015-02-12 23:00:00
100           GDSC_SA    3230    2945 2015-02-12 23:00:00
               SCAN_DATE CELL_ID MASTER_CELL_ID COSMIC_ID CELL_LINE_NAME
99  2015-02-16 23:00:00    4712            198    753608          PC-14
100 2015-02-16 23:00:00    4712            198    753608          PC-14
    SEEDING_DENSITY DRUGSET_ID ASSAY DURATION POSITION     TAG DRUG_ID
99              250        159   Glo        4       99    NC-1    4000
100             250        159   Glo        4      100 L1-D1-S    1003
    CONC INTENSITY
99    NA     38027
100  0.1     23093
```

GDSC raw data definitions

Each row in the raw data represents a single well of a plate. However, there may be more than one row per well if there is more than one tag for that position in the drug set, e.g. this will happen if a well receives a combnination of treatments.

Column_name	Description
RESEARCH_PROJECT	Project name for the dataset
BARCODE	Unique barcode for screening assay plate
SCAN_ID	Unique id for the scan of the plate by the plate reader - fluorescence measurement data. A plate might be scanned more than once but only one `SCAN_ID` will pass internal QC. Therefore there is a one to one correspondence between `BARCODE` and `SCAN_ID` in the published data.
DATE_CREATED	Date that the plate was seeded with cell line.
SCAN_DATE	Date the experiment finished and measurement was taken (scanning).
CELL_ID	Unique GDSC identifier for the cell line expansion seeded on the plate. Each time a cell line is expanded from frozen stocks it is assigned a new `CELL_ID`.
MASTER_CELL_ID	Unique GDSC identifier for the cell line seeded on the plate. A particular cell line will have a single `MASTER_CELL_ID` but can have multiple `CELL_ID`.
COSMIC_ID	Identifier of the cell line in the COSMIC database if available. There is a one to one correspondence between `MASTER_CELL_ID` and `COSMIC_ID`.
CELL_LINE_NAME	Name of the plated cell line. Again this will have a one to one correspondence with `MASTER_CELL_ID`.
SEEDING_DENSITY	Number of cells seeded per well of screening plate. This number is the same for all wells on a plate.
DRUGSET_ID	The set of drugs used to treat the plate and the associated plate layout.
ASSAY	End point assay type used to assess cell viability, e.g., `Glo` is *Promega CellTiter-Glo*.
DURATION	Duration of the assay in days from cell line drug treatment to end point measurement.
POSITION	Plate well position numbered row-wise. 1536 well plates have 48 columns and 384 well plates have 24.
TAG	Label to identify well treatment - see description below. It is possible to have more than one tag per well `POSITION` such that in the raw data files (csv) there may be more than one row per plate well position, e.g., `L12-D1-S + DMSO`.
DRUG_ID	Unique identifier for the drug used for treatment. In the absence of a drug treatment, e.g., a negative control this field will be `NA`.
CONC	Micromolar concentration of the drug id used for treatment. As with `DRUG_ID` this field can be `NA`.
INTENSITY	Fluorescence measurement at the end of the assay. The fluorescence is a result of `ASSAY` and is an indicator of cell viability.

The TAG column

- Drug treated wells will have entries in the `DRUG_ID` and `CONC` fields.
- Experimental drug treatments are referred to as library drugs. Positive controls and reference compounds are also drug treatments.
- In combination drug treatments there will be an anchor drug in addtion to a library drug.
- `-S` indicates a single treatment, `-C` indicates a combination treatment.
- Each library or anchor drug number (`Lx`, `Ax`) will correspond to a particular drug (`DRUG_ID`). However, it is possible that the same `DRUG_ID` will have been assigned to different library or anchor numbers, e.g., to distinguish replicate treatments.
- The relative concentrations of library drug treatments are indicated by the dose (`...-Dx-...`) such that dose `D1` is the maximum concentration and all subsequent doses are dilutions thereof.

Examples of the tags currently in use is given below.

Drug treated wells:

TAG	Description
L1-D1-S	Library drug 1 at dose 1 (maximum concentration) as single agent treatment
L2-D5-S	Library drug 2 alone (combination treatment) at dose 5 (the minimum in a 5 point titration)
A1-C	Anchor drug 1 in a combination
A1-S	Anchor drug 1 alone
R1-D1-S	Reference compound used for comparison between screens

Control wells:

TAG	Description
NC-0	Negative control (no treatment)
NC-1	Negative control (treatment with DMSO)
PC-1	Positive control. No titration of this positive control in the drug set
PC1-D1-S	Positive control as part of a titration.
UN-USED	Excluded from analysis (no cells). Usually wells at the plate edge.
B	Blank (no drug, no cells, just media)
DMSO	Usually used with a drug treatment tag at the same position to indicate back-filling to a required volume.
SC	Cell seeding control with DMSO. A multiple of the cell seeding density used for the rest of the plate.