SUPPLEMENTARY INFORMATION

PharmacoDB: an integrative database for mining in vitro anticancer drug screening studies

Supplementary Methods

Sources for the pharmacogenomic studies

For the Cancer Cell Line Encyclopedia, the data was downloaded from the CCLE portal (`https://portals.broadinstitute.org/ccle`), using the February 24, 2015 update of the pharmacological profiling data. The data from the Genomics of Drug Sensitivity in Cancer project was downloaded from the dedicated portal (`http://www.cancerrxgene.org/`), using release 6 of the data, corresponding to the GDSC1000 update of the dataset. The gCSI dataset was made available with Haverty et al. (Nature 2016) in the compareDrugScreens R package (`http://research-pub.gene.com/gCSI-cellline-data/`), which was fetched July 2016. The CTRP dataset was retrieved from the National Cancer Institute Cancer Target Discovery and Development (CTD^2) data portal, using the v2 release of the data (CTRPv2; `https://ocg.cancer.gov/programs/ctd2/data-portal`). The OHSU GRAY dataset was obtained from the supplementary data of Daemen et al. (Genome Biology 2013).

Fitting drug dose-response curves

The rate-limiting step in the killing of susceptible cancer cells by an anti-cancer drug is assumed to be the binding of a target receptor R to n molecules of a drug D, and the free drug molecules and unbound target are assumed to be in thermodynamic equilibrium with the bound drug-target complex:

$$R + nD \leftrightarrow RD_n$$

By the Law of Mass Action, this equilibrium is characterized by an equilibrium constant K satisfying the equation

$$K = \frac{[RD_n]}{[R][D]^n}$$

where brackets around a variable name denote the concentration of the molecule it represents. It then follows that the fraction f of receptors bound to ligands is given by

$$f = \frac{[RD_n]}{[R] + [RD_n]} = \frac{[D]^n}{[D]^n + K}$$

Since this binding is the rate-limiting step in the killing of cancer cells, the fraction $y(x)$ of susceptible cancer cells killed by a concentration x of the drug is approximately equal to the fraction of unbound receptors $1 - f$:

$$y(x) \approx 1 - f = \frac{1}{1 + (\frac{x}{\sqrt[n]{K}})^n}$$

We define $EC_{50} = \sqrt[n]{K}$ and interpret it as the concentration of the drug needed to have half of the target receptors bound at equilibrium. We also assume that a fraction E_∞ of the cancer cells are not susceptible to the drug at all. This leads to our final equation

$$y(x) = E_\infty + \frac{1 - E_\infty}{1 + (\frac{x}{EC_{50}})^{HS}}$$

where $y(x) = 0$ denotes death of all treated cells, $y(x) = 1$ denotes no effect of the drug dose, EC_{50} is the concentration at which $y(EC_{50}) = \frac{1}{2}$, and HS is a parameter describing the cooperativity of binding. $HS < 1$ denotes negative binding cooperativity, $HS = 1$ denotes noncooperative binding, and $HS > 1$ denotes positive binding cooperativity.

This is the basic mathematical structure that was posited to underlie the dose-response data observed in the study. Consequently, median cellular viability data from all datasets was fit by means of least-squares regression to equations of this type. To ensure robustness of the curve-fitting algorithms,

bounds were placed on the values of each of these parameters. Drugs were assumed not to increase the fitness of malignant cells, so E_∞ was constrained to lie in the interval $[0, 1]$. Drugs were also assumed to have EC_{50} values within $[1\text{pM}, 1\text{M}]$, an interval containing the EC_{50} values reported by Barretina et al. (Nature 2012). Finally, we follow Fallahi et al. (Nat Chem Biol 2013) in allowing HS to lie anywhere in $[0, 4]$.

Barretina et al. (Nature 2012) fit dose-response data to one of three models. In most cases, their model of choice was identical to our own, with the addition of a maximum viability parameter E_0. Their dose response equation then became

$$y = E_\infty + \frac{E_0 - E_\infty}{1 + (\frac{x}{EC_{50}})^{HS}}$$

The inclusion of this parameter makes comparison of dose-response curves problematic. With its inclusion, the viability of the cell line in the absence of any drug becomes

$$y(0) = E_\infty + \frac{E_0 - E_\infty}{1 + (\frac{0}{EC_{50}})^{HS}} = E_\infty + E_0 - E_\infty = E_0$$

As a result, the viability measures of different drug-cell line combinations are normalized differently, and direct comparison of viability predictions from different dose-response curves is no longer appropriate. The IC_{50} values they reported, however, were simply the concentrations at which their fitted curves reached viability reduction of 50% of cellular viability. The end result was a reported IC_{50} value that assumed normalization of viability data to the negative control associated with a curve fitted assuming normalization of viability data to a reference level that was most consistent with the observed data. The IC_{50} values published in the paper's supplementary information thus represented viability reduction by a fraction that varied from cell line to cell line.

In GDSC, the following five-parameter model was used:

$$y = E_\infty + \frac{E_0 - E_\infty}{(1 + (\frac{x}{EC_{50}})^{HS})^S}$$

However, since the E_0 parameter is fixed by controls, their curve can be represented as

$$y = E_\infty + \frac{1 - E_\infty}{(E_\infty(1 + (\frac{x}{EC_{50}})^{HS})^S}$$

This parameter accounts for the presence of an antagonistic binding of the drug, and introduces asymmetry into the theoretical log dose-response curve. The extra parameter, known as the "Schild slope", allows the dose-response curve to be non-monotonic.

While this parameter is well-founded biologically, we chose not to use it in our own dose-response curves. As only medians of technical replicates are available for CCLE, using a 4-parameter model would have increased our susceptibility to overfitting noise in the sparse dose-response curves. Furthermore, we only rarely observed the non-monotonicity that necessitates the inclusion of a Schild slope parameter in a very small fraction of dose-response curves. For these reasons, we ultimately chose to use our simpler 3-parameter model to compare the dose-response curves from the GDSC and CCLE datasets.

Gene Drug Associations

The association between molecular features and response to a given compound was computed using the `drugSensitivitySig` function in *PharmacoGx*. The function models the association using a linear regression model adjusted for tissue source of the cell line:

$$Y = \beta_0 + \beta_i G_i + \beta_t T + \beta_b B$$

where Y denotes the sensitivity variable, G_i, T and B denote the expression of gene i, the tissue source and the experimental batch respectively, and βs are the regression coefficients. The strength of gene-compound association is quantified by β_i, above and beyond the relationship between drug sensitivity and tissue source. The variables Y and G are scaled (standard deviation equals to 1) to estimate standardized coefficients from the linear model and increase the comparability of the estimates between genes and compounds. Significance of the gene-drug association is estimated by the statistical significance of β_i (two-sided t test). The p-values in PharmacoDB are left uncorrected as the number of associations for a given gene and compound change with the integration of datasets and new molecular types, therefore changing the number of hypotheses tested.

Code for Tissue Specific Case Study

```
library(PharmacoGx)

CCLE <- downloadPSet("CCLE")

all.tissues <- drugSensitivitySig(CCLE, "rna",
                  features = "ENSG00000141736",
                  drugs = "lapatinib",
                  sensitivity.measure="AAC")["ENSG00000141736","lapatinib
                      ",]

print("Strength of associations across all tissues:")
print(all.tissues["estimate"])

by.tissue <- do.call(cbind,tapply(cellNames(CCLE), cellInfo(CCLE)$tissueid
    , function(x) {
    tissue.only <- subsetTo(CCLE, cells=x)
    if(length(x) < 5){
        warning(paste0("Tissue ", cellInfo(CCLE)[x[1],"tissueid"], "
            contained not enough cell lines to compute statistics."))
        return(rep(NA_real_, 8))
    }
    if(!"lapatinib" %in% drugNames(tissue.only)){
        warning(paste0("lapatinib was not tested in tissue: ", cellInfo(
            CCLE)[x[1],"tissueid"]))
        return(rep(NA_real_, 8))
    }
    return(drugSensitivitySig(tissue.only, "rna",
                              features="ENSG00000141736",
                              drugs = "lapatinib",
                              sensitivity.measure="AAC")["ENSG00000141736"
                                  ,"lapatinib",])
}))

print("Strength of associations by tissue type:")
print(by.tissue["estimate",order(by.tissue["estimate",], decreasing=TRUE)
    ])
```

Application Programming Interface (API)

Overview

The base URL used for all queries is https://api.pharmacodb.com/v1/. The end âĂIJ/v1âĂİ corresponds to the version of the API being queried. This version can be changed to any of the API versions that are, or will be, released by PharmacoDB. Whenever a newer version is released, both the API and the MySQL database of the previous version are frozen and open for use at that version URL. Furthermore, breaking changes will only be introduced in a newer version. Hence, no application that integrates or uses PharmacoDB data will break, or be affected by breaking changes.

The API currently supports a wide range of search queries, based on the resource type being queried. Resource types include: *cell lines*, *tissues*, *drugs*, *datasets*, *experiments*, *intersections*, and *stats*. All available endpoints for each resource type are documented in the API's main repository on github at `https://github.com/bhklab/PharmacoDB`.

Request/Response Format

All requests use HTTP GET method, and the base URL used for making requests is `https://api.pharmacodb.com/v1/`. Valid endpoints should be concatenated to the base URL when making a request. All valid endpoints are listed on github, as well as the **API Reference** section below. No API keys, or tokens, are needed in order to make an API call.

To demonstrate, the following is a sample request using the **curl** command:

```
$ curl "https://api.pharmacodb.com/v1/cell_lines"
```

The above request queries a list of cell lines found in PharmacoDB. All resource requests return a paginated list by default. This means that if PharmacoDB contans N number of cell lines, only the first K items are returned at initial request, and the rest N - K items can be queried by changing the page value. The *page* and *per_page* options are available for modifying paginated lists. For example, the above request can be modified to get the third page, whereby each page contains 20 items, as follows:

```
$ curl "https://api.pharmacodb.com/v1/cell_lines?page=3&per_page=20"
```

If, instead, all cell lines need to be queried in a single request, the request will look as follows:

```
$ curl "https://api.pharmacodb.com/v1/cell_lines?all=true"
```

Setting the *all=true* option overrides default pagination, and returns a list of all items found in a resource type of interest.

To select a single item in a resource type, an ID is used as a parameter, as follows:

```
$ curl "https://api.pharmacodb.com/v1/cell_lines/{id}"
```

In the above, *id* corresponds to either a cell line ID, or a cell line name. By default, the API assumes the request is using a cell line ID. For example, making a request to:

```
$ curl "https://api.pharmacodb.com/v1/cell_lines/1"
```

retrieves the cell line whose ID = 1. In order to query a cell line by its name instead of ID, one can use the *type* option, as follows:

```
$ curl "https://api.pharmacodb.com/v1/cell_lines/mcf7?type=name"
```

Metadata, such as pagination information and links, are all included in response header by default. This metadata can be included in response body by using the *include=metadata* option. For example, making a request to:

```
$ curl "https://api.pharmacodb.com/v1/cell_lines?include=metadata"
```

retrieves a list of cell line items, whereby pagination information and links metadata are included in response body.

All results are indented, or pretty printed, by default. Users can customize results and turn off indentation by using the *indent=false* option, as follows:

```
$ curl "https://api.pharmacodb.com/v1/cell_lines?indent=false"
```

All endpoint options are listed below in the **API Reference** section.

Errors

All valid responses return with HTTP status code 200 (Status OK). Any other status code in response signals an error. There are three main error types defined in the API.

1. **400 (Bad Request)**: This error type is returned when no routers match the request URL. This can happen when the requested endpoint is not one that is defined in the API. The API documentation lists all available endpoints.
2. **404 (Not Found)**: A 404 error is returned whenever a resource item is not found in database.
3. **500 (Internal Server Error)**: This is returned whenever a general error occurs in the API itself instead of user query. Users should contact us if met with 500 Internal Server Error responses.

API Reference

Endpoint	Reference
/cell_lines	**Description**: Retrieves a list of all cell lines in PharmacoDB. **Optional**: *page*, *per_page*, *include*, *all*, *indent* **Example**: `https://api.pharmacodb.com/v1/cell_lines?page=2&per_page=5&indent=false`
/cell_lines/{id}	**Description**: Retrieves a single cell line. **Optional**: *indent*, *type* **Example**: `https://api.pharmacodb.com/v1/cell_lines/mcf7?type=name`
[other]	**More endpoints for cell lines**: - /cell_lines/{id}/drugs
/tissues	**Description**: Retrieves a list of all tissues. **Optional**: *page*, *per_page*, *include*, *all*, *indent* **Example**: `https://api.pharmacodb.com/v1/tissues?all=true`
/tissues/{id}	**Description**: Retrieves a single tissue. **Optional**: *indent*, *type* **Example**: `https://api.pharmacodb.com/v1/tissues/breast?indent=false&type=name`
[other]	**More endpoints for tissues**: - /tissues/{id}/cell_lines - /tissues/{id}/drugs
/drugs	**Description**: Retrieves a list of all drugs. **Optional**: *page*, *per_page*, *include*, *all*, *indent* **Example**: `https://api.pharmacodb.com/v1/drugs`
/drugs/{id}	**Description**: Retrieves a single drug. **Optional**: *indent*, *type* **Example**: `https://api.pharmacodb.com/v1/drugs/1`
[other]	**More endpoints for drugs**: - /drugs/{id}/cell_lines - /drugs/{id}/tissues
/datasets	**Description**: Retrieves a list of all datasets. **Optional**: *page*, *per_page*, *include*, *all*, *indent* **Example**: `https://api.pharmacodb.com/v1/datasets?page=1&per_page=2`
/datasets/{id}	**Description**: Retrieves a single dataset. **Optional**: *indent*, *type* **Example**: `https://api.pharmacodb.com/v1/datasets/2`
[other]	**More endpoints for datasets**: - /datasets/{id}/cell_lines - /datasets/{id}/tissues - /datasets/{id}/drugs
/experiments	**Description**: Retrieves a list of all experiments in PharmacoDB. **Optional**: *page*, *per_page*, *include*, *indent* **Note**: The maximum number of experiments that can be returned in a single request is set to 1000. **Example**: `https://api.pharmacodb.com/v1/experiments?page=2&per_page=50`
/experiments/{id}	**Description**: Retrieves a single experiment. **Optional**: *indent*, *type* **Example**: `https://api.pharmacodb.com/v1/experiments/2`

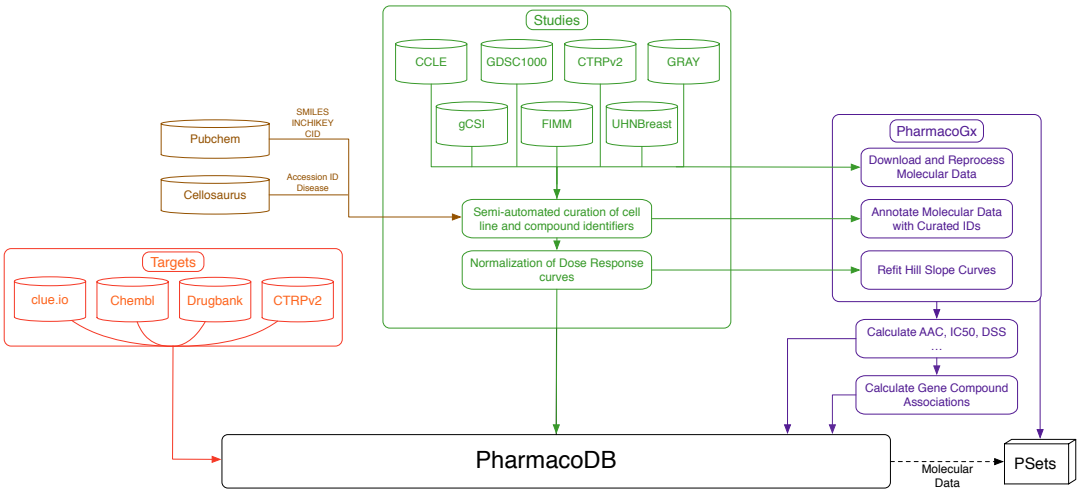
Endpoint	Reference
[Intersections]	**Available endpoints for intersections**: - /intersections/1/{cell_id}/{drug_id} - /intersections/1/{cell_id}/{dataset_id}
[Stats]	**Available endpoints for stats**: - /stats/tissues/cell_lines - /stats/datasets/cell_lines - /stats/datasets/cell_lines/{id}/drugs - /stats/datasets/tissues - /stats/datasets/tissues/{id}/cell_lines - /stats/datasets/tissues/{id}/drugs - /stats/datasets/drugs - /stats/datasets/drugs/{id}/cell_lines - /stats/datasets/drugs/{id}/tissues - /stats/datasets/experiments

Acronyms

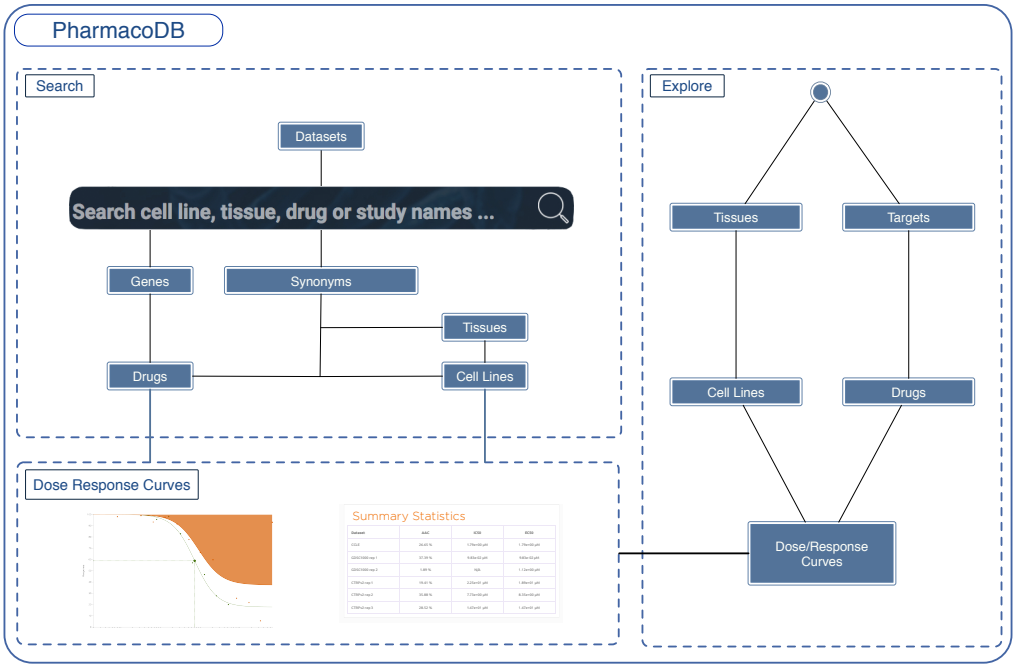
AAC	Area above the dose response curve
CCLE	The Cancer Cell Line Encyclopedia initiated by the Broad Institute of MIT and Harvard
CMAP	Connectivity Map by the Broad Institute
CTRP	Cancer Therapeutic Response Portal
DSS	Drug sensitivity score
EC_{50}	Dose at which 50% of the maximum response is observed
E_{max}	Maximum theoretical inhibition
FIMM	Institute for Molecular Medicine Finland cell viability screen
gCSI	Genentech Cell Screening Initiative
GDSC	The Cancer Genome Project initiated by the Wellcome Trust Sanger Institute
IC_{50}	Concentration at which the drug inhibited 50% of the maximum cellular growth
InchiKey	International Chemical Identifier
OHSU	Oregon Health and Science University
pIC_{50}	$-\log_{10}(IC_{50})$
SMILES	Simplified molecular-input line-entry system
UHN	University Health Network

Supplementary Figures

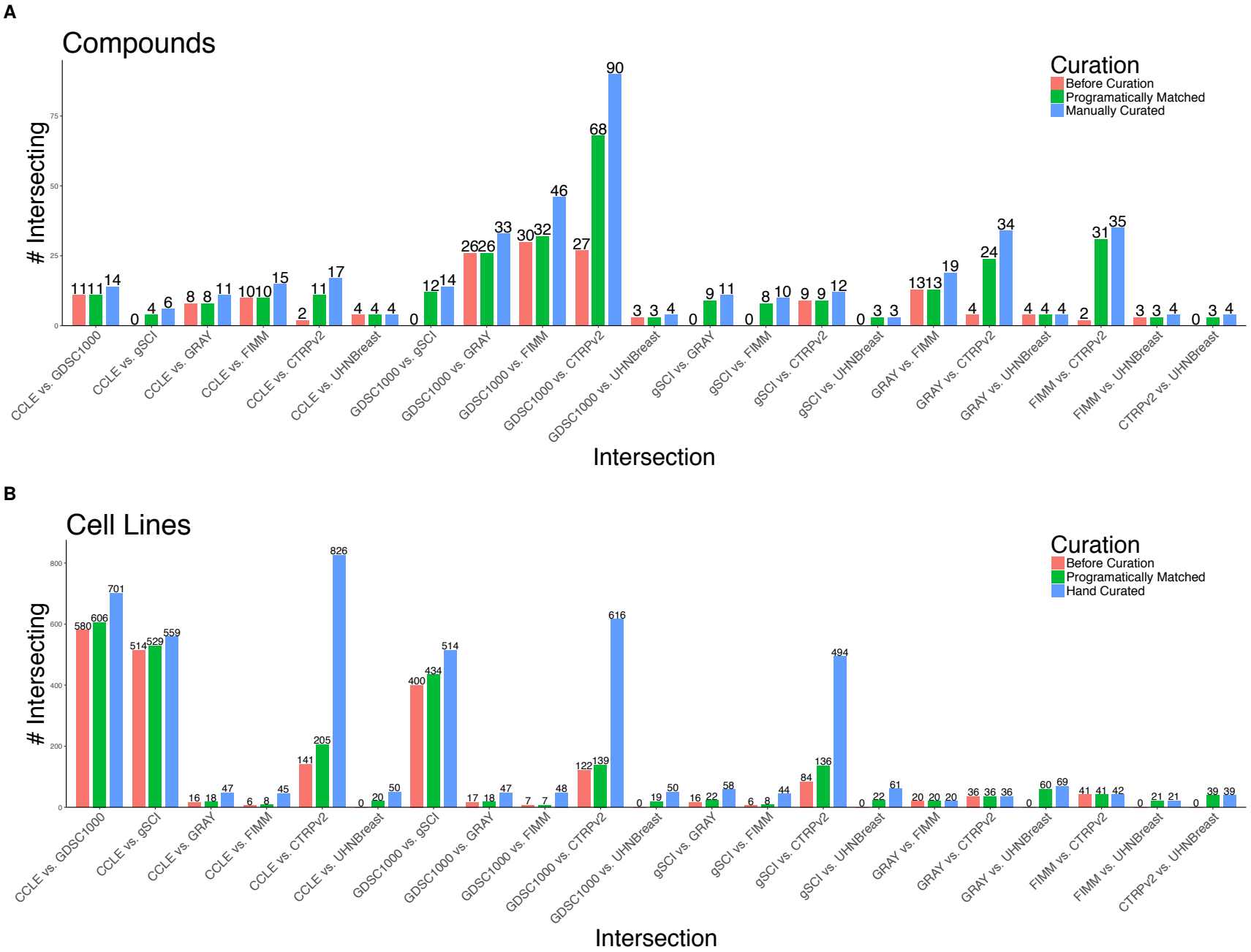
A



B

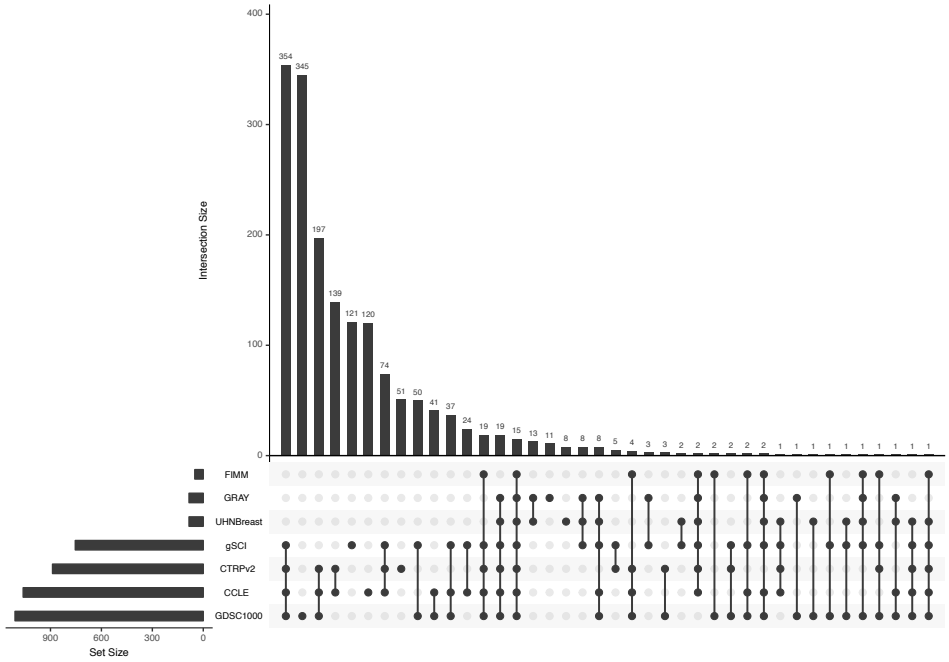


Supplementary Figure 1: Flow diagrams describing (**A**) the external resources used to populate PharmacoDB; and (**B**) the "Search" and "Explore" interfaces to query PharmacoDB.

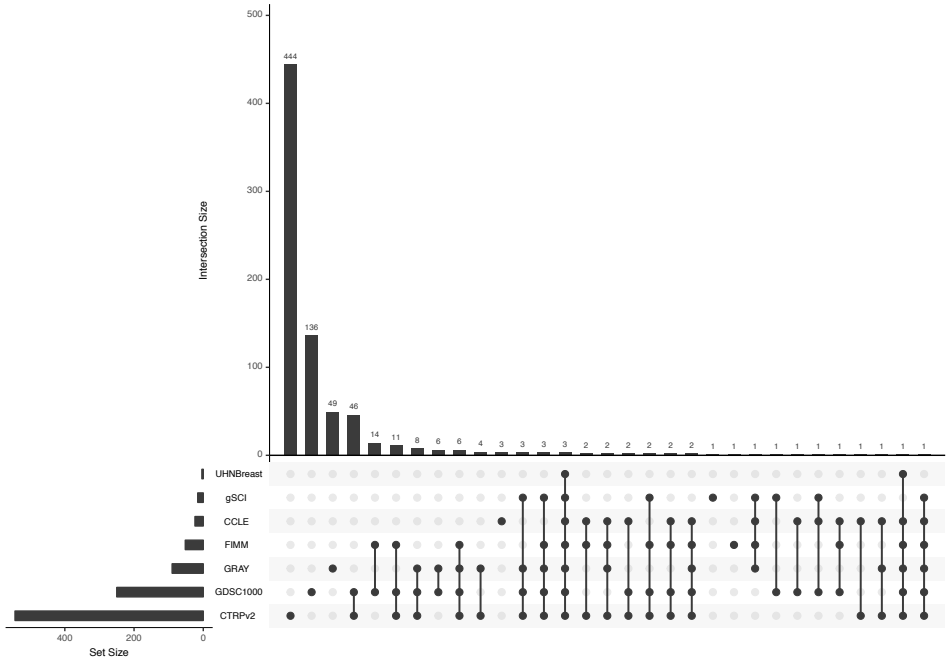


Supplementary Figure 2: The difference in overlap between datasets before curation, after automatically removing differences in capitalization and white space and manually reviewing matches within edit distance of 2, and after undergoing full manual curation for (**A**) cell lines and (**B**) compounds.

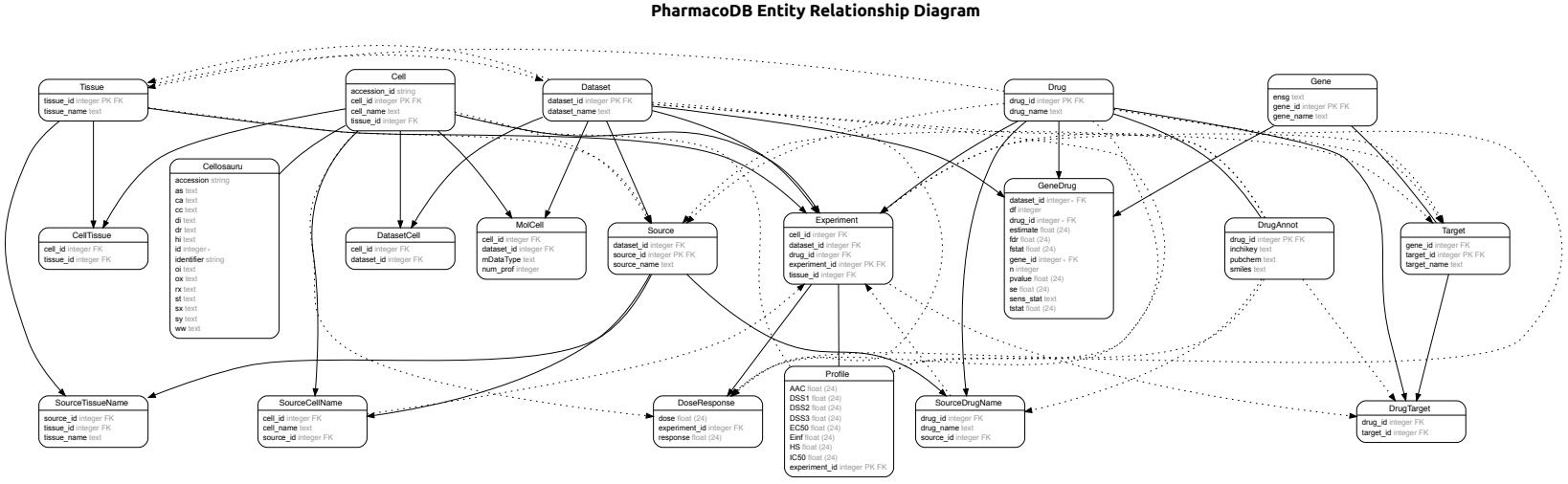
A



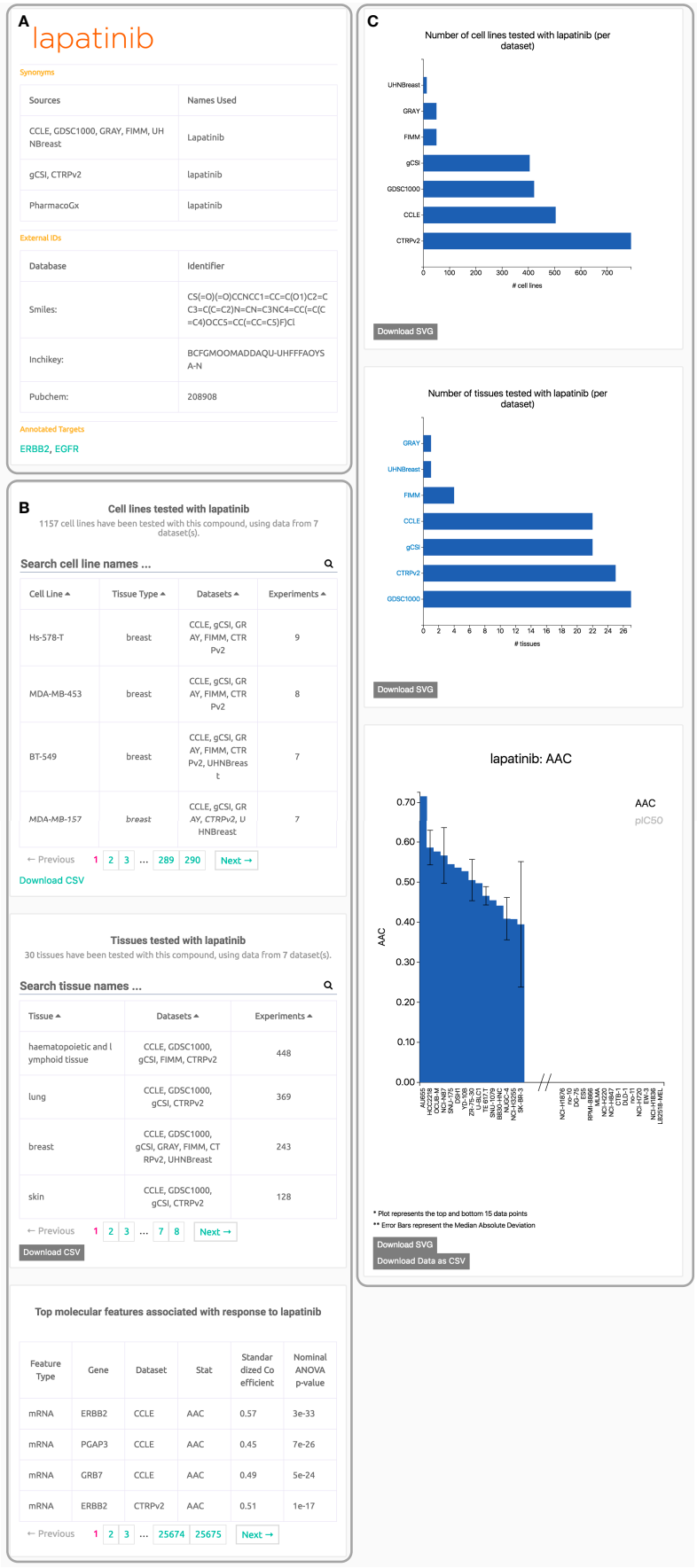
B



Supplementary Figure 3: Intersection between datasets included in PharmacoDB for (**A**) cell lines and (**B**) compounds, after curation of cell line and compound identifiers, respectively.



Supplementary Figure 4: Structure of the database at the level of the object models and their relationships.



Supplementary Figure 5: An example profile page from PharmacoDB for a compound. Panel (**A**) is the information card for the compound (lapatinib). Panel (**B**) contains tables listing the available data profiles for this cell. This includes a table of the genetic associations found across all datasets in PharmacoDB. Panel (**C**) contains summary plots about the drug screening performed in each dataset and the waterfall plots of the cell response to treatment with compounds.