Minimal information for Chemosensitivity assays (MICHA): A next-generation pipeline to enable the FAIRification of drug screening experiments

Ziaurrehman Tanoli[1,9], Jehad Aldahdooh[1,9], Farhan Alam[1], Yinyin Wang[1], Umair Seemab[1], Maddalena Fratelli[2], Petr Pavlis[3], Marian Hajduch[3], Florence Bietrix[4], Philip Gribbon[5], Andrea Zaliani[5], Matthew D. Hall[6], Min Shen[6], Kyle Brimacombe[6], Evgeny Kulesskiy[7], Jani Saarela[7], Krister Wennerberg[8], Markus Vähä-Koskela[7] and Jing Tang[1,*]

Affiliation: 1) Research Program in Systems Oncology, Faculty of medicine, University of Helsinki, Finland; 2) Mario Negri Institute for Pharmacological Research, Italy; 3) Institute of Molecular and Translational Medicine, Czech; 4) European infrastructure for translational medicine; 5) Fraunhofer Institute for Molecular Biology and Applied Ecology, Germany; 6) National Center for Advancing Translational Sciences, U.S.A; 7) Institute for Molecular Medicine Finland, University of Helsinki, Finland; 8) Biotech Research & Innovation Centre (BRIC), University of Copenhagen, Denmark; 9) Shared first authors

***Correspondence:** jing.tang@helsinki.fi

ABSTRACT

Chemosensitivity assays are commonly used for preclinical drug discovery and clinical trial optimization. However, data from independent assays are often discordant, largely attributed to uncharacterized variation in the experimental materials and protocols. Spurred by the annotation of minimum information (MI) for ensuring data reproducibility, we report here the launching of MICHA (Minimal Information for Chemosensitivity Assays), accessed via https://micha-protocol.org. Distinguished from existing MI efforts that are often lack of support from data integration tools, MICHA can automatically extract publicly available information to facilitate the assay annotation including: 1) compounds, 2) samples, 3) reagents, and 4) data processing methods. For example, MICHA provides an integrative web server and database to obtain compound annotation including chemical structures, targets and disease indications. In addition, the annotation of cell line samples, assay protocols and literature references can be greatly eased by retrieving manually curated catalogues. Once the annotation is complete, MICHA can export a report that conforms to the FAIR principle (Findable, Accessible, Interoperable and Reusable) of drug screening studies. To consolidate the utility of MICHA, we provide *FAIRified* protocols from several major cancer drug screening studies, as well as recently conducted COVID-19 studies. With the integrative webserver and database, we envisage a wider adoption of the MICHA strategy to foster a community-driven effort to improve the open access of drug sensitivity assays.

KEYWORDS

Drug discovery, drug sensitivity assays, data integration tools, FAIR research data.

BACKGROUND

Drug sensitivity or chemo sensitivity assay is an important tool to measure cellular response to drug perturbation, which has been increasingly used for preclinical drug discovery and clinical trial optimization. However, poor inter- and intra-laboratory reproducibility has been reported when comparing batches that differ at assay materials and methods[1-3]. Central to improving the data reproducibility is the standardization of assay ontologies, with which the research community may annotate minimal information (MI) about critical components of a typical drug sensitivity experiment. An existing implementation is MIABE (minimum information about a bioactive entity), which provides guidelines for the annotation of compound bioactivities[4]. However, MIABE mainly was designed for drug-target profiling assays and thus it is not directly applicable for drug sensitivity assays. Furthermore, like many other similar MI efforts in omics assays, MIABE provided limited data integration tools to enable its implementation coherently.

The solution we present here, MICHA (minimal information for chemosensitivity assays), provides an integrative pipeline to facilitate the annotation of four major components of a drug sensitivity assay, including 1) compounds tested, 2) samples, 3) reagents, and 4) data processing references. Using the platform of MICHA, we aim to increase acceptance and adoption of the principles of FAIR (Findable, Accessible, Interoperable and Reusable), by making the assay annotation as smoothly as possible with the help of data integration tools and databases.

MATERIALS AND METHODS

Work flow

Using MICHA, users can upload their chemicals, samples and experimental design information (**Figure 1**). To start, users need to upload the names and InChiKeys of the compound, after which MICHA will automatically extract primary and secondary target information, physio-chemical properties and disease indications for these compounds. Users may continue filling in the other experimental details, such as sample (cell lines or patient-derived organoids) information and assay conditions. For cell lines, only the names of the cell lines are required, as the other information will be retrieved automatically from internal databases. For annotating assay conditions, we derived a consensus on the minimal information that is needed, including assay format, detection technology, end point mode of action, experimental medium, plate type, cell density, time for treatment, dilution fold, vehicle of compound, dispensation method and volume per well. The definition of these

terminologies are explained in **Supplementary File 1** as well as in the 'Glossary' tab at the MICHA website. Furthermore, users are directed to a web form to report data processing information, including minimum and maximum concentrations of the compounds, publication references and drug response metric names such as IC_{50} or AUC. Finally, a tabular report can be generated according to the user's input combined with the augmented annotation from public resources. In addition, users may browse annotated protocols including major cancer and COVID-19 drug screening studies.

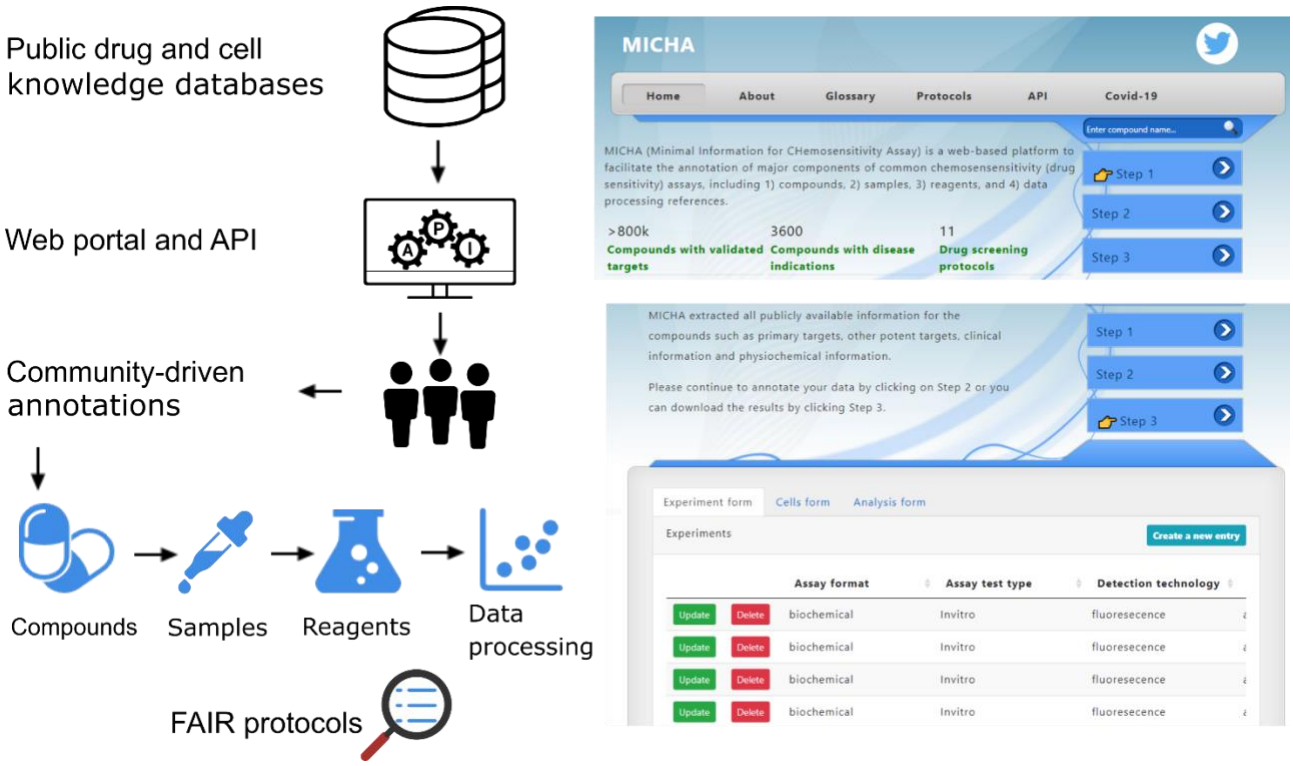

***Figure 1**: User interface and work flow of MICHA. Users start with compound annotation by uploading a list of compounds with names and InChiKeys, and obtain the compound pharmacological and physio-chemical properties via an integrative webserver and database. Users may continue to annotate reagents/samples, experiments and data processing methods, and generate a report to include all the information in a tabular format. MICHA also harbors the annotated drug sensitivity screening protocols for cancer and COVID-19.*

Data integration tools

Three types of datasets are retrievable via data integration tools in MICHA:

1) ***Drug-target data***

Compound target profiles were integrated from the most comprehensive drug target databases including DTC[5,6], BindingDB[7], ChEMBL[8], GtopDB[9], DGiDB[10] and DrugBank[11]. The first four databases (DTC, BindingDB, ChEMBL and GtopDB) contain quantitative bioactivity data whereas DGiDB and DrugBank contains unary drug target information. We focused on the primary and secondary targets of a compound, defined as those displaying binding affinities <= 1000 nM from the bioactivity databases, or those that are recorded in unary

databases. We integrated drug targets for 277K chemicals from DTC, 513K from ChEMBL, 258K from Binding DB, 4.8K from GtopDB, 7.6K from DGiDB and 6.8K from DrugBank. Furthermore, we merged overlapping targets across these databases in order to avoid duplications, resulting in high-quality target profiles for >800K chemicals. Such a data integration provided one of the most comprehensive compounds collection along with their potent primary and secondary targets. All the drug-target information can be retrieved by the MICHA annotation pipeline or programmatically using the API (application programming interface).

2) ***Compound properties, cell line and assay information***

Compound physiochemical properties and structures for 1.9 million compounds were obtained from the ChEMBL database. Furthermore, we integrated disease indications and clinical phase information for 3600 drugs from the DTC database. This information together with the drug-target profiles will be retrieved for user-uploaded compound list. When users annotate the cell lines, the majority of cell line information can be retrieved automatically from Cellosaurus, which is a comprehensive knowledge database on cell lines[12]. For assay annotation, commonly used techniques will be provided for users to choose from to ease the burden of manual editing.

3) ***FAIRified protocols***

A prime objective for MICHA is to provide a pipeline for the FAIRification of drug sensitivity assays, such that these established protocols can be well documented with enhanced visibility to the research community. To initiate such an effort, we have FAIRified drug screening protocols from major cancer studies including GDSC (345 compounds and 987 cell lines)[13], CCLE (24 compounds and 504 cell lines)[14] and CTRPv2 (203 compounds and 242 cell lines) [15]. Furthermore, we have provided drug sensitivity screening protocols extracted from six recent COVID-19 antiviral studies (5,525 compounds and 2 cell lines) [16-21]. On the other hand, we have provided an example of protocols adopted at the institution level (528 compounds and 4 cell lines utilized at the high throughput drug-screening unit at the Institute for Molecular Medicine Finland, University of Helsinki). These FAIRified protocols can be freely obtained at http://micha-protocol.org/protocols/. With more protocols annotated via MICHA, the research community shall be better informed on the variations on the experimental condition across different studies and institutions.

ADDED VALUES BY MICHA

Comprehensive drug-target profiles

MICHA has integrated various databases, which harbor drug-target data at different levels, ranging from quantitative bioactivity values to unary drug target hits. For instance, DrugBank, GtopDB and DGiDB are mainly focused on approved compounds with putative target information, while ChEMBL, BindingDB and DTC are inclusive of bioactivity values for more versatile investigational and pre-clinical chemicals. In MICHA, we have tried to improve target coverage across the druggable genome by integrating non-overlapping data points from the latest releases of these databases. As shown in **Figure 2**, the average number of targets for 2993 approved drugs in MICHA is 7.33, as compared to that from ChEMBL (5.5), DGiDB (4.71), DrugBank (3.56), BindingDB (2.74) and GtopDB (0.96). Similarly, for 1992 investigational compounds (defined as those in clinical trials), the average number of targets per chemical is higher in MICHA as compared to other databases. The improved target coverage in MICHA shall open up new applications for drug discovery researchers for training their drug-target prediction models [22, 23] as well as providing more insights on the mode of actions.

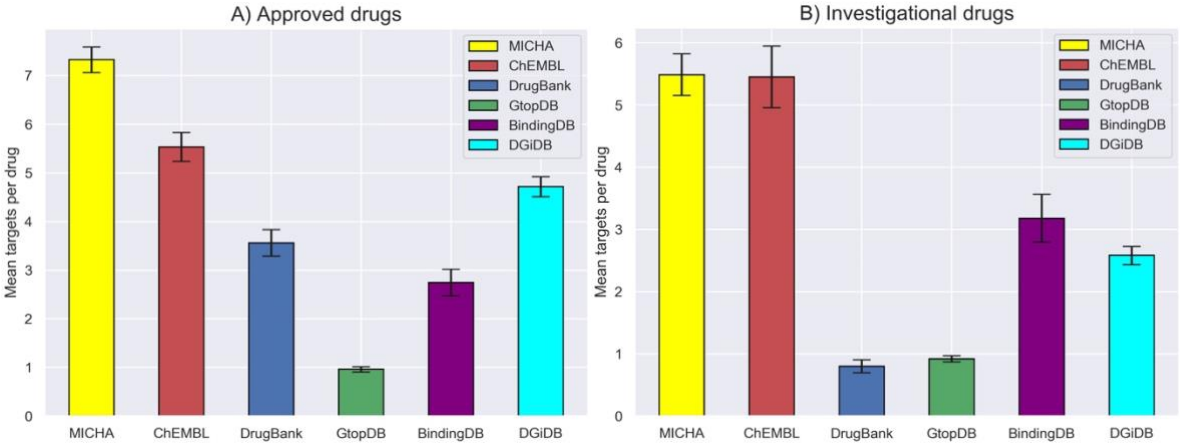

Figure 2: *Average number of targets for approved and investigational drugs in multiple databases.*

FAIRification of drug sensitivity screening protocols

We have already FAIRified three major cancer drug screening protocols including CCLE, GDSC and CTRPv2. These studies share similar objectives of linking genetic features of cancer cell lines to small-molecule sensitivity to accelerate drug discovery. **Table 1** shows a comparison of the major components in the protocols.

Table 1: Comparison between protocols of CCLE, GDSC and CTRP. NA indicates that the information is unavailable.

	CCLE	***GDSC***	***CTRPv2***
Plate types	1536	384, 96	384-Opaque white

Experimental mediums	NA	DMEM, RPMI	ALPHAMEM, DMEM, DMEM:F12 (1:1), EMEM, F12K, HamsF10, HamsF12, IMDM, L15, McCoy's5A, MCDB105:MDB 199 (1:1), MEM, RPMI, RPMI:F12 (1:1), Waymouth's MB 752/1, William's E
Detection technology	Luminescence	Fluorescence	Luminescence
Cell density	250	NA	500, 1000, 2000, 3000, 5000, 10000
Treatment time (h)	72-84	72	72
Analysis metrics	IC_{50}, EC_{50}	IC_{50}	IC_{50}
Compounds	24	345	203
Number of cell lines	504	987	242

Both GDSC and CTRPv2 used common experimental plate type i.e. 384 wells; whereas CCLE data was tested on 1536 well plates. In GDSC, two different experimental mediums including DMEM and RPMI were tested for the 987 cancer cell lines, whereas the CTRPv2 cell lines were tested for many more different mediums as listed in **Table 1**. In contrast, we could not find experimental medium information for CCLE. On the other hand, both CCLE and CTRPv2 have used Cell-Titer-Glo (Promega), a luminescence-based assay to measure levels of ATP as a surrogate method to cell viability, whereas GDSC has used based nucleic acid staining syto60 (Invitrogen) for adherent cells and resazurin (Sigma) for suspension cells. Treatment time is standardized using hours and all the three screening studies have used at least 72 hours of treatment, after which the IC_{50} or EC_{50} concentrations were determined from the dose response curves.

Figure 3 shows the overlapping chemicals and cell lines tested across CCLE, GDSC and CTRPv2 studies, after excluding those chemicals for which proper chemical names or identifiers were missing in order to preserve high quality data in MICHA. Only two chemicals are shared across the three studies including selumetinib and tanespimycin (**Figure 3A**). Selumetinib (AZD6244) is a MEK (kinase) inhibitor used for treating neurofibromatosis type I in children[24]; where as tanespimycin is a Hsp90 inhibitor[25] which has been studied for the treatment of leukemia or solid tumors, especially kidney tumors. In contrast, more overlap was found for the cell lines, with 112 cell lines in common across the three studies (**Figure 3B**).

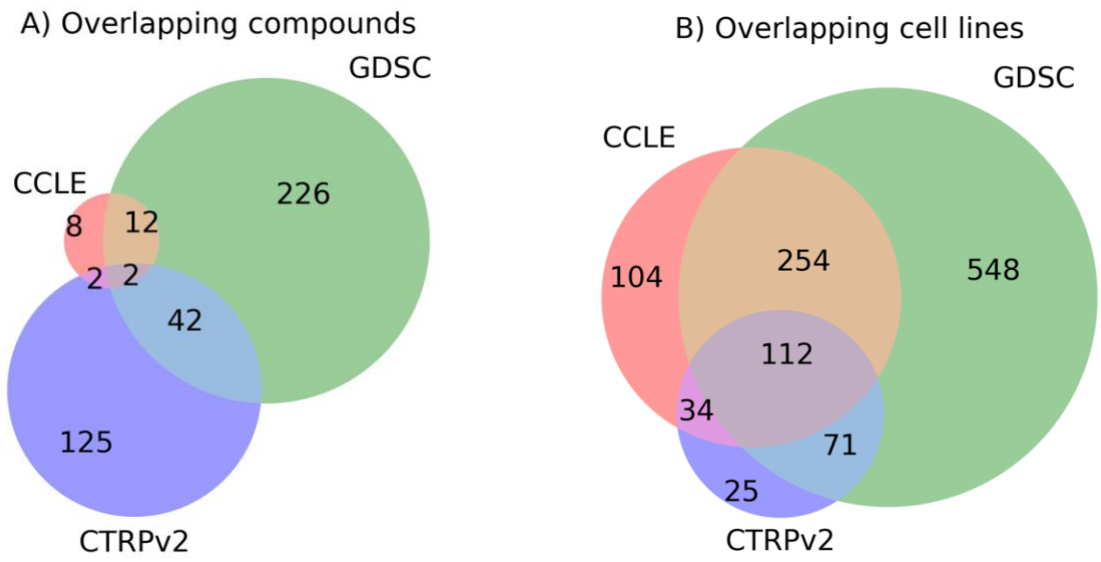

Figure 3*: Overlapping compounds and cell lines between CCLE, GDSC and CTRPv2.*

MICHA has not only FAIRified cancer related drug screening studies but also annotated six recent studies on COVID-19, a virus that causes ongoing pandemic with limited drug treatment options. From these studies, we have identified 5,318 chemicals tested across two cell lines including Vero E6 and Caco-2. The annotations of these compounds, cell lines, as well as the experimental information and data analysis methods can be easily retrieved at http://micha-protocol.org/covid19. With the help of MICHA platform, all these COVID-19 and cancer related drug screening protocols are freely accessible to the users (Findable). Moreover, these protocols can be accessed programmatically using MICHA API available at: http://micha-protocol.org/api (Accessible). *In silico* models can programmatically access MICHA to obtain compound information such as protein targets and physiochemical properties and use this information for novel drug target predictions (Interoperable). MICHA drug screening protocols as defined in aforementioned sections can be used as a reference for the experimental design of future drug screening studies, as well as serve as a source of information to evaluate the data reproducibility (Reusable). Moreover, the MICHA source code is available at GitHub for reuse of the application towards a tailoring need for individual drug screening centers (https://github.com/JehadAldahdooh/MICHA).

CONCLUSION

Chemosensitivity assay screening has been increasingly utilized for preclinical drug discovery and clinical trial optimization. However, chemosensitivity assays often lack sufficient annotation to make the data FAIR, which has become a limiting factor for supporting its clinical translation. To improve the assay annotation, web portals that facilitate information retrieval from different assay components are critically needed. To address this issue, we have recently launched MICHA (Minimal information for Chemosensitivity assays) as a web server for the annotation of chemosensitivity screens that covers critical information including 1) compounds 2) samples 3) reagent protocols and finally, 4) data processing methods.

Comprehensive drug-target profiles were deposited to the MICHA database for more than 800K compounds. These high-quality pharmacological data shall help improve the understanding on the mode of actions not only for approved drugs but also for investigational and preclinical compounds. Furthermore, the target profiles at the druggable genome scale provide more information on the polypharmacological effects, which might lead to new opportunities for drug repositioning[26]. To facilitate the data retrieval, the API in MICHA is highly optimized such that it can return target profiles for thousands of chemicals within seconds.

With the MICHA web portal, we have FAIRified major drug sensitivity screening protocols from five cancer studies and six recent COVID-19 studies, serving as the first instances of the catalogue. Comparing these manually curated assay protocols should allow a more systematic analysis of data reproducibility. With the FAIR-compliant data resources and tools to deliver content standards and ontology services, MICHA will ensure the characterization of critical assay components, allow the FAIRification and cataloguing of drug sensitivity studies, and support the downstream analysis towards clinical translation. We invite the drug discovery community to use MICHA for annotating their drug sensitivity assays to improve the knowledge sharing, which shall ultimately lead to a bigger impact in translational medicine.

Acknowledgement

EU H2020 (EOSC-LIFE, No. 824087); ERC (DrugComb, No. 716063); Academy of Finland (No. 317680). CSC-IT Center for Science for providing computing resources.

REFERENCES

1. Hatzis, C.; Bedard, P. L.; Birkbak, N. J.; Beck, A. H.; Aerts, H. J.; Stern, D. F.; Shi, L.; Clarke, R.; Quackenbush, J.; Haibe-Kains, B., Enhancing reproducibility in cancer drug screening: how do we move forward? *Cancer Res.* **2014,** *74* (15), 4016-4023.

2. Haverty, P. M.; Lin, E.; Tan, J.; Yu, Y.; Lam, B.; Lianoglou, S.; Neve, R. M.; Martin, S.; Settleman, J.; Yauch, R. L., Reproducible pharmacogenomic profiling of cancer cell line panels. *Nature* **2016,** *533* (7603), 333-337.
3. Larsson, P.; Engqvist, H.; Biermann, J.; Rönnerman, E. W.; Forssell-Aronsson, E.; Kovács, A.; Karlsson, P.; Helou, K.; Parris, T. Z., Optimization of cell viability assays to improve replicability and reproducibility of cancer drug sensitivity screens. *Sci. Rep.* **2020,** *10* (1), 1-12.
4. Orchard, S.; Al-Lazikani, B.; Bryant, S.; Clark, D.; Calder, E.; Dix, I.; Engkvist, O.; Forster, M.; Gaulton, A.; Gilson, M., Minimum information about a bioactive entity (MIABE). *Nature reviews Drug discovery* **2011,** *10* (9), 661-669.
5. Tanoli, Z.; Alam, Z.; Vähä-Koskela, M.; Ravikumar, B.; Malyutina, A.; Jaiswal, A.; Tang, J.; Wennerberg, K.; Aittokallio, T., Drug Target Commons 2.0: a community platform for systematic analysis of drug–target interaction profiles. *Database* **2018,** *2018*.
6. Tang, J.; Ravikumar, B.; Alam, Z.; Rebane, A.; Vähä-Koskela, M.; Peddinti, G.; van Adrichem, A. J.; Wakkinen, J.; Jaiswal, A.; Karjalainen, E., Drug target commons: a community effort to build a consensus knowledge base for drug-target interactions. *Cell chemical biology* **2018,** *25* (2), 224-229. e2.
7. Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J., BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* **2016,** *44* (D1), D1045-D1053.
8. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E., The ChEMBL database in 2017. *Nucleic Acids Res.* **2017,** *45* (D1), D945-D954.
9. Alexander, S. P.; Kelly, E.; Marrion, N. V.; Peters, J. A.; Faccenda, E.; Harding, S. D.; Pawson, A. J.; Sharman, J. L.; Southan, C.; Buneman, O. P., The concise guide to pharmacology 2017/18: overview. *Br. J. Pharmacol.* **2017,** *174*, S1-S16.
10. Wagner, A. H.; Coffman, A. C.; Ainscough, B. J.; Spies, N. C.; Skidmore, Z. L.; Campbell, K. M.; Krysiak, K.; Pan, D.; McMichael, J. F.; Eldred, J. M., DGIdb 2.0: mining clinically relevant drug–gene interactions. *Nucleic Acids Res.* **2016,** *44* (D1), D1036-D1044.
11. Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J., DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **2006,** *34* (suppl_1), D668-D672.
12. Bairoch, A., The cellosaurus, a cell-line knowledge resource. *Journal of biomolecular techniques: JBT* **2018,** *29* (2), 25.
13. Iorio, F.; Knijnenburg, T. A.; Vis, D. J.; Bignell, G. R.; Menden, M. P.; Schubert, M.; Aben, N.; Gonçalves, E.; Barthorpe, S.; Lightfoot, H., A landscape of pharmacogenomic interactions in cancer. *Cell* **2016,** *166* (3), 740-754.
14. Barretina, J.; Caponigro, G.; Stransky, N.; Venkatesan, K.; Margolin, A. A.; Kim, S.; Wilson, C. J.; Lehár, J.; Kryukov, G. V.; Sonkin, D., The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **2012,** *483* (7391), 603-607.
15. Seashore-Ludlow, B.; Rees, M. G.; Cheah, J. H.; Cokol, M.; Price, E. V.; Coletti, M. E.; Jones, V.; Bodycombe, N. E.; Soule, C. K.; Gould, J., Harnessing connectivity in a large-scale small-molecule sensitivity dataset. *Cancer Discov.* **2015,** *5* (11), 1210-1223.
16. Stuart Weston, C. M. C., Rob Haupt, James Logue, Krystal Matthews, Matthew B. Frieman, Broad anti-coronaviral activity of FDA approved drugs against SARS-CoV-2 in vitro and SARS-CoV in vivo. **2020**.
17. Gordon, D. E.; Jang, G. M.; Bouhaddou, M.; Xu, J.; Obernier, K.; White, K. M.; O'Meara, M. J.; Rezelj, V. V.; Guo, J. Z.; Swaney, D. L., A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* **2020,** *pp*, 1-13.

18. Touret, F.; Gilles, M.; Barral, K.; Nougairède, A.; van Helden, J.; Decroly, E.; de Lamballerie, X.; Coutard, B., In vitro screening of a FDA approved chemical library reveals potential inhibitors of SARS-CoV-2 replication. *Sci. Rep.* **2020,** *10* (1), 1-8.
19. Jeon, S.; Ko, M.; Lee, J.; Choi, I.; Byun, S. Y.; Park, S.; Shum, D.; Kim, S., Identification of antiviral drug candidates against SARS-CoV-2 from FDA-approved drugs. *Antimicrobial Agents Chemotherapy* **2020**.
20. Brimacombe, K. R.; Zhao, T.; Eastman, R. T.; Hu, X.; Wang, K.; Backus, M.; Baljinnyam, B.; Chen, C. Z.; Chen, L.; Eicher, T., An OpenData portal to share COVID-19 drug repurposing data in real time. *bioRxiv* **2020**.
21. Ellinger, B.; Bojkova, D.; Zaliani, A.; Cinatl, J.; Claussen, C.; Westhaus, S.; Reinshagen, J.; Kuzikov, M.; Wolf, M.; Geisslinger, G., Identification of inhibitors of SARS-CoV-2 in-vitro cellular toxicity in human (Caco-2) cells using a large scale drug repurposing collection. **2020**.
22. Cichonska, A.; Pahikkala, T.; Szedmak, S.; Julkunen, H.; Airola, A.; Heinonen, M.; Aittokallio, T.; Rousu, J., Learning with multiple pairwise kernels for drug bioactivity prediction. *Bioinformatics* **2018,** *34* (13), i509-i518.
23. Cichonska, A.; Ravikumar, B.; Parri, E.; Timonen, S.; Pahikkala, T.; Airola, A.; Wennerberg, K.; Rousu, J.; Aittokallio, T., Computational-experimental approach to drug-target interaction mapping: a case study on kinase inhibitors. *PLoS Comput. Biol.* **2017,** *13* (8), e1005678.
24. Gross, A. M.; Wolters, P. L.; Dombi, E.; Baldwin, A.; Whitcomb, P.; Fisher, M. J.; Weiss, B.; Kim, A.; Bornhorst, M.; Shah, A. C., Selumetinib in children with inoperable plexiform neurofibromas. *N. Engl. J. Med.* **2020,** *382* (15), 1430-1442.
25. Neckers, L.; Workman, P., Hsp90 molecular chaperone inhibitors: are we there yet? *Clin. Cancer Res.* **2012,** *18* (1), 64-76.
26. Tanoli, Z.; Seemab, U.; Scherer, A.; Wennerberg, K.; Tang, J.; Vähä-Koskela, M., Exploration of databases and methods supporting drug repurposing: a comprehensive survey. *Brief. Bioinform.* **2020**.