Workflow for Defining Reference Chemicals to Assess Performance of *In Vitro* Assays

Anita Simha[3], Richard S. Judson[1], Russell S. Thomas[1], Nancy Baker[2], Xia Meng Howey[3], Carmen Marable[3], Nicole C. Kleinstreuer[4], Keith A. Houck[1]

[1]U.S. Environmental Protection Agency, Office of Research and Development | [2]Leidos, Inc. | [3]Oak Ridge Associated Universities | [4]National Toxicology Program

Anita Simha | simha.anita@epa.gov | 919-541-2766

Abstract

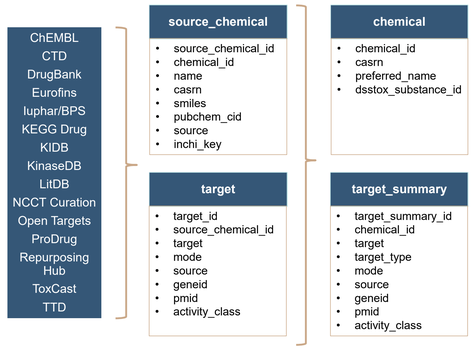
OBJECTIVE: Develop a semi-automated workflow for identifying and annotating reference chemicals to validate *in vitro* assay data.

BACKGROUND: Use of reference chemicals is key for validating high-throughput *in vitro* assay data used in predictive toxicology, but developing reference chemical lists has historically only occurred on a small scale due to context specificity and resource limitations.

METHODS: We extracted information from curated and non-curated open-source databases on molecular target, chemical, and mode of action into **RefChemDB**. We also compared data from EPA's ToxCast program to results from the literature. To determine support, we tallied independent reports of each unique chemical-target-mode combination. To contextualize support from different data sources, we manually validated a subset of the data that is associated with PubMed IDs and determine precision rates across hand curators.

RESULTS: We compiled a database with 1,234,580 unique chemical-target-mode combinations. Performance of ToxCast bioassays strongly correlated with the level of support for a chemical-target-mode combination. We hand-curated data for 54 molecular targets, with precision rates of 82.7% from curated sources and 39.5% from automated literature extraction, informing our workflow recommendation. A list of candidate reference chemicals was created by selecting chemical-target-mode combinations with a minimum support level of 5 records.

Database Structure

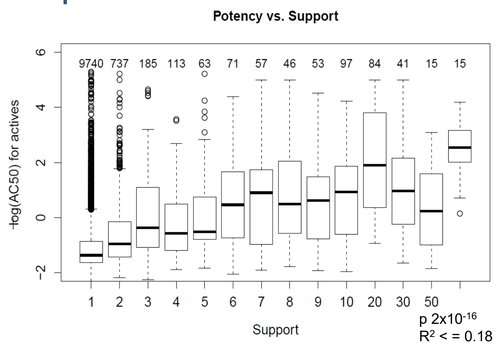


Source	Chemicals	Targets	Chemical-Target-Mode combinations	PMIDs
ChEMBL	28832	2238	310984	11520
ChEMBL Drug	1187	738	4099	0
CTD	2317	7904	25606	5280
DrugBank	1630	1169	3623	6274
Eurofins Biochemical	206	570	925	0
Eurofins Functional	211	239	706	0
Iuphar BPS	1860	941	5081	0
KEGG Drug	661	263	1201	0
KIDB	535	450	6532	0
KinaseDB	133	168	676	1
LitDB	2654	88	8348	27909
Open Targets	1031	820	3973	0
Prodrug	41	33	41	1
Repurposing Hub	2279	2172	10209	0
ToxCast	9136	343	852470	0
TTD	3916	1575	11557	0
Web Curation	3940	1059	5617	0
Total	**37301**	**11055**	**1234580**	**49883**

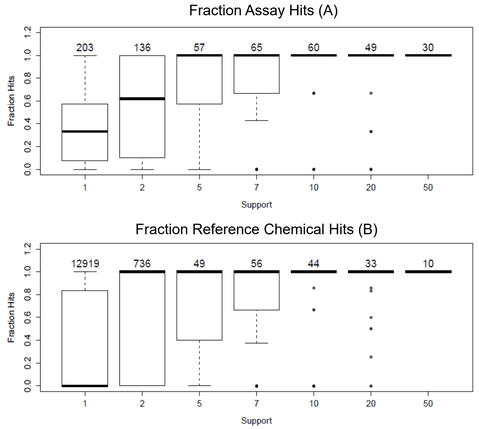
RefChemDB structure and contents.

Results

Comparison with ToxCast



Correlation between level of support in RefChemDB and potency in ToxCast.



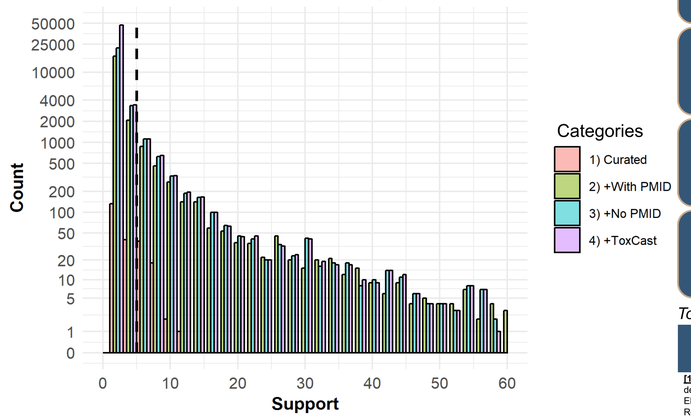
(A): Fraction of reference chemicals active in an assay relative to support.
(B): Fraction of assays that a chemical was active in relative to support.

Candidate Reference Chemicals

Target Symbol	Target Name	Chemicals	Support
CA2	Carbonic Anhydrase 2	106	2453
CA1	Carbonic Anhydrase 1	105	1981
ESR1	Estrogen Receptor 1	85	1371
DRD2	Dopamine Receptor D2	81	952
AR	Androgen Receptor	63	750
ESR2	Estrogen Receptor 2	41	664
PTGS2	Prostaglandin-Endoperoxide Synthase 2	46	646
OPRM1	Opioid Receptor Mu 1	48	638
CA9	Carbonic Anhydrase 9	23	512
PPARA	Peroxisome Proliferator-Activated Receptor Alpha	27	511
PPARG	Peroxisome Proliferator-Activated Receptor Gamma	26	486
NR3C1	Nuclear Receptor Subfamily 3 Group C Member 1	37	482
HTR2A	5-Hydroxytrtyptamine Receptor 2A (Serotonin Receptor 2A)	43	476
ADRB2	Beta-2 Adrenergic Receptor	44	472
ACHE	Acetylcholinesterase	28	470
SLC6A4	Solute Carrier Family 6 Member 4	40	423
ABCB1	ATP Binding Cassette Subfamily B Member 1	35	418
KCNH2	Potassium Voltage-Gated Channel Subfamily H Member 2	48	412
HRH1	Histamine Receptor H1	45	399
HDAC1	Histone Deacetylase 1	12	387

Top 20 most-supported targets with tally of chemical interactions observed in literature with total support.

Support Across 4 Categories



Frequency of support for a particular target-chemical-mode combination. Support values ≥2 and ≤60 are shown for ease of visualization.

Hand Curation

Validation rate between curators/sources and precision rate among curators (above). Equality of proportions test results among sources (below).

Source 1	Source 2	p-value
ChEMBL	CTD	0.574
ChEMBL	DrugBank	0.0337*
CTD	DrugBank	0.7761
ChEMBL	LitDB	<2.2e-16***
CTD	LitDB	<2.2e-16***
DrugBank	LitDB	<2.2e-16***

Summary

- **Query RefChemDB**
  - Return target, chemical, mode, and source information from the database
- **Evaluate reference candidate lists**
  - Identify chemicals with support > 5
  - Identify chemicals in highly accurate sources (DrugBank, ChEMBL)
- **Validate data**
  - Use PubMed IDs where possible to confirm assay data and metadata with two reviewers
  - Finalize with subject matter expert review
- **Deep dive into literature**
  - Use literature search tools like EPA's PubMed Abstract Sifter to find more definitive target/chemical interaction information

To access RefChemDB files, please see Judson et al. 2018 in ALTEX

Resources

[1] Jaworska, J., N. Nikolova-Jeliazkova and T. Aldenberg (2005). "QSAR applicabilty domain estimation by projection of the training set descriptor space: a review." Altern Lab Anim **33**(5): 445-459. **[2]** NICEATM / ICCVAM. (2011). "NICEATM DRAFT ED BRD: BG1Luc ER TA Test Method – Section 3.0." **[3]** Mansouri, K., A. Abdelaziz, A. Rybacka, A. et al (2016). "CERAPP: Collaborative Estrogen Receptor Activity Prediction Project." Environ Health Perspect 124(7): 1023-1033. **[4]** Tilkkainen, P., L. Bellis, Y. Light and L. Franke (2013). "Estimating error rates in bioactivity databases." J Chem Inf Model **53**(10): 2499-2505. **[5]** Casey, Chang, et al. (2018). "Evaluation and Optimization of Pharmakokinetic Models for *in Vitro* to *in Vivo* Extrapolation of Estrogenic Activity for Environmental Chemicals." Environ. Health Perspect. 126(9): 097001-14. **[6]** Baker, N., Knudsen, T., Williams, A. (2017). "Abstract Sifter: a comprehensive front-end system to PubMed." F1000Res. 6: 2164.