Future
Drug Discovery

Research Article

How polypharmacologic is each chemogenomics library?

Eric Ni[1,2,3], Eehjoe Kwon[1], Lauren M. Young[2], Klara Felsovalyi[2], Jennifer Fuller[2] & Timothy Cardozo*[,1]

[1]NYU Langone Health, Department of Biochemistry & Molecular Pharmacology, New York, NY 10016, USA
[2]Genecentrix Inc, New York, NY 10014, USA
[3]Yale University, Department of Computational Biology & Bioinformatics, New Haven, CT 06520, USA
*Author for correspondence: Timothy.Cardozo@nyumc.org

Aim: High-throughput phenotypic screens have emerged as a promising avenue for small-molecule drug discovery. The challenge faced in high-throughput phenotypic screens is target deconvolution once a small molecule hit is identified. Chemogenomics libraries have emerged as an important tool for meeting this challenge. Here, we investigate their target-specificity by deriving a 'polypharmacology index' for broad chemogenomics screening libraries. **Methods:** All known targets of all the compounds in each library were plotted as a histogram and fitted to a Boltzmann distribution, whose linearized slope is indicative of the overall polypharmacology of the library. **Results & conclusion:** Comparison of libraries clearly distinguished the most target-specific library, which might be assumed to be more useful for target deconvolution in a phenotypic screen.

First draft submitted: 28 October 2019; Accepted for publication: 16 December 2019; Published online: 5 February 2020

Keywords: chemogenomics • polypharmacology • target deconvolution

High-throughput phenotypic screening (pHTS) is defined as direct application of perturbagens (usually drug-like small molecule compounds) to complex biological systems that exhibit complex phenotypes (usually cells, but commonly organoids and animals as well) [1]. This approach is based on the prioritization of drug candidate cellular bioactivity over drug mechanism of action (MoA). As phenotypic screening takes place in a physiologically relevant environment of cells or in whole organisms, the hits appear to have a greater probability of success at later stages in drug development. Specifically, new medical entities may derive more efficiently from phenotype screens than traditional, high-throughput, biochemical or biophysical target-based screening (tHTS) [2]. In addition, pHTS can link drug-like compounds to bioactivities that are the product of an ensemble of proteins and pathways that cannot be modeled biochemically. However, identifying the molecular targets of active hits from pHTS, also known as target deconvolution, is required to understand the phenotype mechanism and to further optimize active compounds.

tHTS for drug discovery tests thousands to millions of compounds against a single target protein. A common outcome of tHTS is that the compound exhibits insufficient *in vivo* efficacy in animal models of the target disease. This frequently occurs even when excellent biophysical or biochemical activity of a compound against the target is attained and when such activity is optimized by medicinal chemistry and/or structure-based optimization. Conversely, leads obtained from pHTS appear to have a higher rate of success at the *in vivo* pharmacodynamics stage of drug development [2]. One reason for this may be underestimation of the plasticity and chemical and metabolic complexity of biological systems, which has led to attempts to measure the 'distance' between bioassays and the true *in vivo* scenario [3].

Compounds optimized against single targets have been repurposed for use in pHTS screens [4–6], based on the assumption that knowledge of the target of each compound confers automatic target deconvolution upon pHTS. The National Institutes of Health's (NIH; MD, USA) Mechanism Interrogation PlatE (MIPE) array and Novartis' (Basel, Switzerland) MoA Box are two examples of large libraries of such assumed target-specific compounds – also termed chemogenomics libraries [7]. However, many compounds emerging from tHTS are known to interact with

newlands
press

10.4155/fdd-2019-0032 © 2020 Timothy Cardozo eISSN 2631-3316

multiple targets (polypharmacology), with most drug molecules interacting with six known molecular targets on average, even after optimization [8]. Polypharmacology and the use of chemogenomics libraries in phenotypic screens to enhance target deconvolution are therefore opposing concepts, but the former has not been widely explored in the context of the latter (although polypharmacology of individual drugs and compounds in the context of the drug development funnel has previously been investigated [9]). Indeed, the only study to date on this question limited its analysis to the data-rich space of kinase inhibitors and found that all available kinase chemogenomics libraries were poorly target annotated [10]. Moret *et al.* used their findings to develop an improved chemogenomics library covering the druggable genome. Here, we develop a quantitative polypharmacology index to better annotate chemogenomics libraries and compare them to this first rationally designed chemogenomics library.

Methods

Chemical libraries

Compound libraries were obtained from publicly available sources. The libraries were chosen based on the suitability of their data, public availability and analysis in recent papers and reviews on chemical genetics. With regard to data suitability, the ones we chose are the only ones that are public, well known and also are annotated with proper chemical identifiers (ChEMBL ID, DrugBank ID, PubChem ID or CAS numbers) and not simply names or internal identifiers. DrugBank [11] was also added even though it does not have all the criteria because it is a broad and general library that includes every drug so we can compare chemogenomic libraries to it. The Spectrum Collection from Microsource Discovery Systems (CT, USA) contains 1761 bioactive compounds for use in HTS- or target-specific assays. The NIH's MIPE [12] library is comprised of 1912 small molecule probes, all of which have a known mechanism of action, contains 9700 compounds, including approved, biotech and experimental drugs, which do not necessarily have drug targets annotated. The Laboratory of Systems Pharmacology – Method of Action (LSP-MoA) library is an optimized chemical library that optimally targets the liganded kinome. CAS numbers and PubChem CIDs from libraries were converted to the Simplified Molecular Input Line Entry System (SMILES) using an ICM script for further processing (Molsoft, LLC, CA, USA). Compounds were cross-registered via their canonical SMILES strings, which preserve stereochemistry information and other variations, such as salts.

Target identification

Target annotations were enumerated as previously described for generating historeceptomic scores [13]. Briefly, *in vitro* binding data was obtained from ChEMBL [14] in the form of Ki and IC_{50} values, or from DrugBank as affinities, for each compound in each library and then filtered for redundancy. Each compound query included compounds related by 0.99 Tanimoto similarity, so that salts, isomers, – among others, of the compound were included in the query. Tanimoto similarity coefficients were calculated in Python, using tool RDkit, which generates molecular fingerprints from chemical data in the form of a SMILES string, and then compares these fingerprints to calculate the Tanimoto similarity coefficient or 'distance'. The number of recorded molecular targets for each compound was recorded and the histograms shown in supplementary Figure 1 were generated using MATLAB from these counts. Notably, since the affinity of the drug for the target likely determines whether it is a true biological target, with nanomolar affinities representing significant targets and micromolar affinities being ambiguous, we assigned target status to any drug–receptor interaction that had a true measured affinity less than the upper limit of the assay. Drug–target interactions with recorded affinities at the upper limit of the assay were assumed to be negative.

Data analysis

The number of hits for each drug in each library was counted and a histogram was created. The histogram values were sorted in descending order and transformed into natural log values using MATLAB's Curve Fitting Suite, which were then used to find the slope of the linearized distribution. The slope is the PP_{index}. MATLAB also solves for the coefficients of an exponential curve, or more simply the log of that curve, that minimize deviations from observed data points (ordinary least squares). All curves have an R square value of above 0.96 for a Boltzmann distribution, indicating goodness of fit.

Optimized libraries for better polypharmacology were created by sequentially eliminating highly promiscuous compounds from the base library individually, while prioritizing high target coverage and optimal PP_{index} with the remaining compounds.

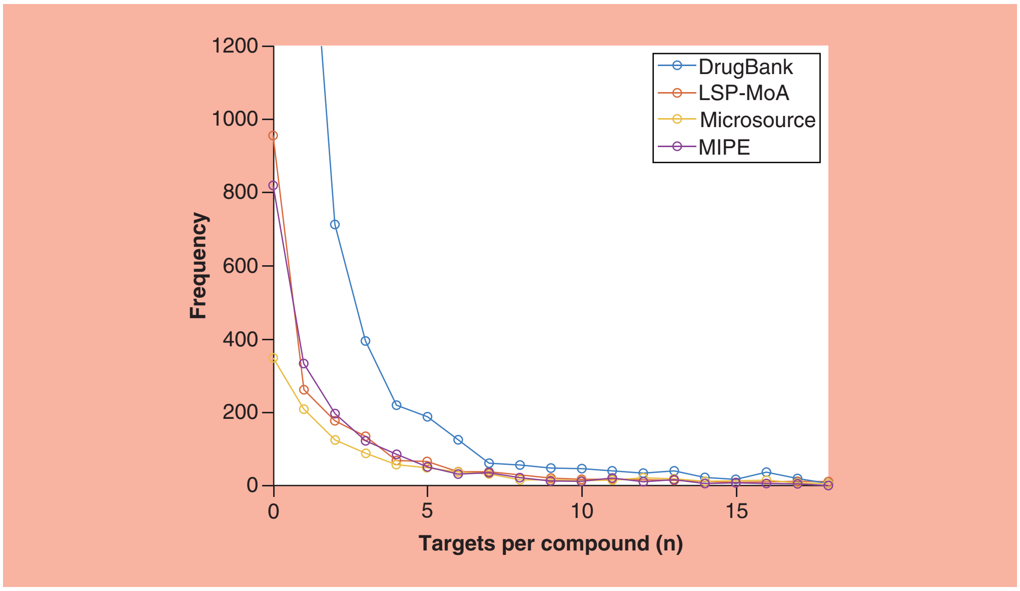


Figure 1. Sample target output for one compound in the analysis demonstrating no target-specific annotation.
LSP-MoA: The Laboratory of Systems Pharmacology–Method of Action; MIPE: Mechanism Interrogation PlatE.

Table 1. Absolute values of slopes of linearized distributions presented in Supplementary Figure 1.

Database	All	Without 0	Without 1+0
DrugBank	0.9594	0.7669	0.4721
MIPE	0.7102	0.4508	0.3847
Microsource spectrum	0.4325	0.3512	0.2586
LSP-MoA	0.9751	0.3458	0.3154
DrugBank approved	0.6807	0.3492	0.3079

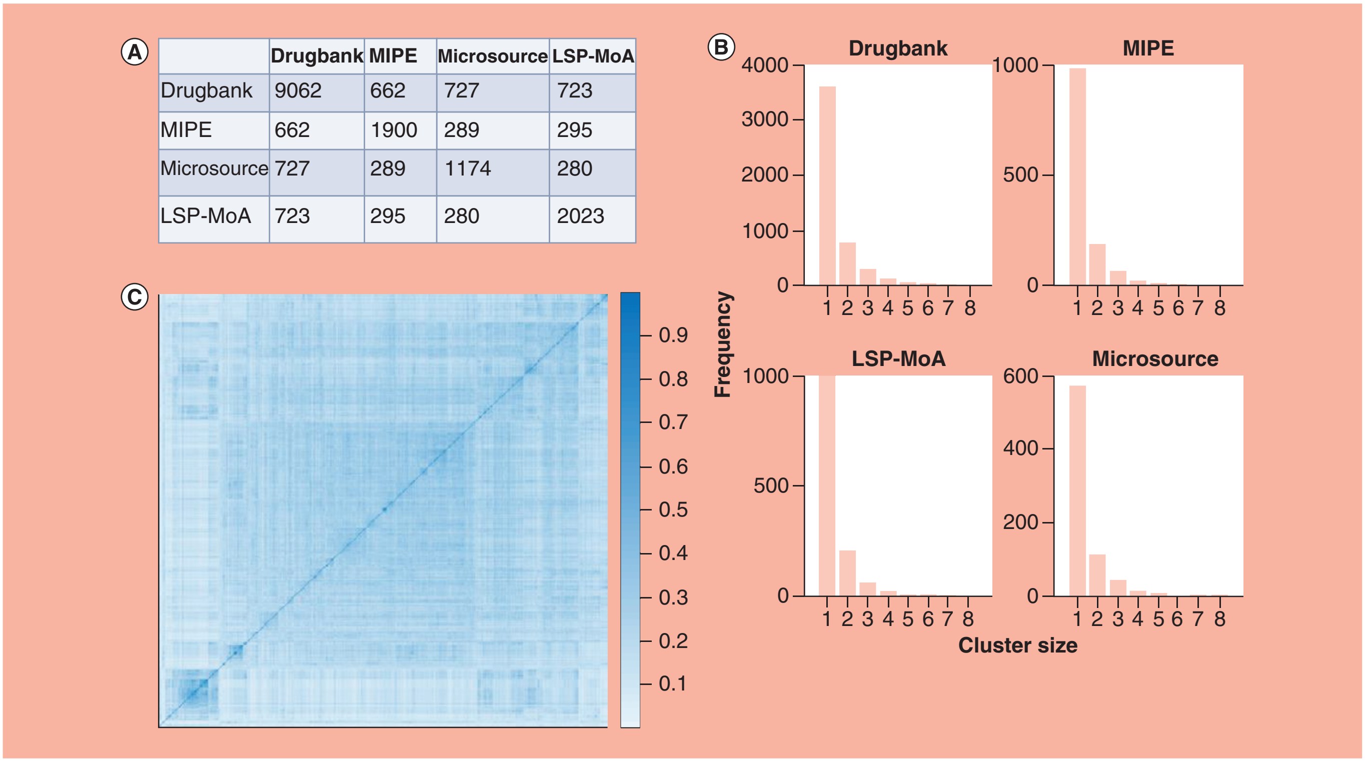
LSP-MoA: The Laboratory of Systems Pharmacology–Method of Action; MIPE: Mechanism Interrogation PlatE.

Results

We compared the following targeted compound libraries: Microsource spectrum, MIPE 4.0 and LSP-MoA libraries to a nontargeted compound library, the DrugBank library. In addition, we also compared a subset of the DrugBank library containing only approved drugs. Histograms of the number of targets per compound exhibited Boltzmann-like distributions. Interestingly, the bin of compounds with no annotated target was the single largest category or subset of the compounds in each library (Figure 1).

The distributions for the LSP-MoA and MIPE 4.0 libraries appear to exhibit an enhanced shoulder as compared with the Microsource library, and both appear to have a fewer number of compounds with a single target as compared with DrugBank. The distributions thus suggest *a priori* that DrugBank is less polypharmacologic than the other three libraries. The DrugBank approved subset also shows this enhanced shoulder, thus higher polypharmacology, compared with its parent library. As a quantitative measure of this observation, linear transformation of the distribution using natural log of the distribution gives a slope for the shoulder of the distribution that is a single number (PP_{index}), which could be representative of the polypharmacology of the library, with larger numbers (slopes closer to a vertical line) being representative of more target-specific libraries and smaller numbers (slopes closer to a horizontal line) being more and more polypharmacologic. The PP_{index} values for all four libraries are shown in Table 1.

While DrugBank superficially appears to be far more target-specific, this is due to its larger size and data sparsity with many compounds identified in the literature as having only one target but having not been screened against any other targets. Accordingly, we linearized the distributions in the absence of the 0-target and 1-target bins of the distribution to reduce this bias. Indeed, the PP_{index} was dramatically reduced, but still showed improved target specificity over the other libraries (Table 1). For ranking and optimizing the diversity of focused small molecule



	Drugbank	MIPE	Microsource	LSP-MoA
Drugbank	9062	662	727	723
MIPE	662	1900	289	295
Microsource	727	289	1174	280
LSP-MoA	723	295	280	2023

Figure 2. Comparison of chemical composition and similarity across drug libraries. (A) Table showing inter-library compound overlap, the number of identical compounds between libraries. **(B)** Frequencies of cluster size when clustered to Tanimoto distance <0.3. **(C)** Chemical similarity against self, for the MIPE library, units in Tanimoto similarity. Other libraries show similar patterns, but are not shown here.

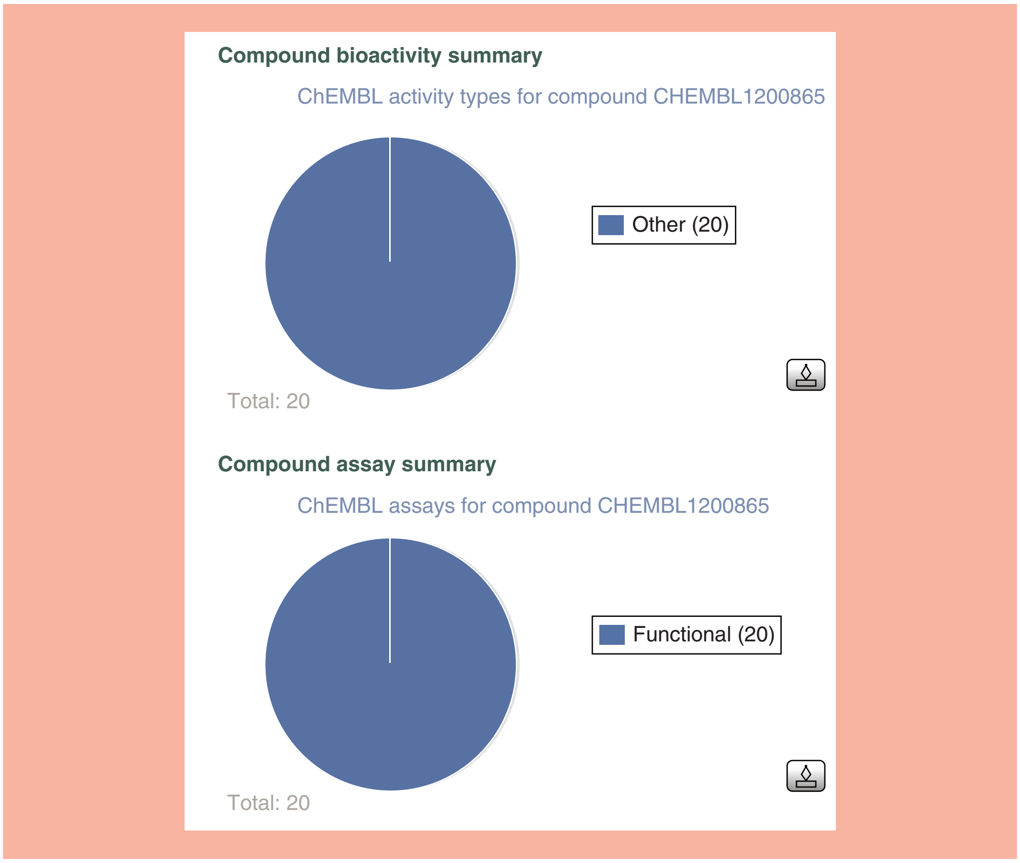
LSP-MoA: The Laboratory of Systems Pharmacology–Method of Action; MIPE: Mechanism Interrogation PlatE.

libraries, the composition of two chemical libraries can be compared based on aggregate structural similarity of their resident compounds. We show the results of this analysis in Figure 2. By generating a phylogenetic tree and setting Tanimoto distance to <0.3, we find that the distribution of the frequencies for cluster size is nearly identical for all libraries (Figure 2B). Plotting chemical similarity shows a similar trend as well, in that these libraries generally all have a high amount of diversity (Figure 2C).

A major use of chemical libraries that are less polypharmacologic is in phenotypic screens, wherein active compounds themselves automatically deconvolute the target if their annotated target is accurate and specific. If one were to run phenotypic screens using highly promiscuous compounds, target deconvolution becomes more difficult as the number of possible targets for an active compound skyrockets. Making better libraries without these problematic compounds can be helpful in this regard. Accordingly, we attempted to optimize the libraries and reduce polypharmacology while maintaining high target coverage by removing the most promiscuous compounds one at a time. Interestingly, this results in a near-linear relationship between the number of compounds and the targets they cover for the LSP-MoA, Microsource and MIPE libraries (Figure 3), while the DrugBank library has a slight exponential curve. This could be explained by the former three being intended as target-specific libraries, therefore biasing their compounds toward a particular set of targets, resulting in highly promiscuous drugs hitting the same targets as the more specific ones. If every drug were to have a unique set of targets, one would expect an exponential relationship between the number of compounds and their targets, which is a trend that the DrugBank library approaches (Figure 3).

Discussion

We developed a single, simple, intuitive, quantitative metric, the PP_{index} that characterizes the polypharmacology of chemogenomics libraries, thus enabling easy comparison between them. This metric represents the log curve of a histogram for the number of drug targets per compound, meaning that higher numbers indicate less polypharma-



Compound bioactivity summary

ChEMBL activity types for compound CHEMBL1200865

Compound assay summary

ChEMBL assays for compound CHEMBL1200865

Figure 3. Compound libraries optimized for polypharmacology by removing highly promiscuous compounds systematically according to how many targets they hit.

cology. Surprisingly, the index reveals that the DrugBank may be more target-specific than previously appreciated; however, this conclusion is supported by the fact that this same conclusion was previously reached for individual marketed drugs [9]. Thus, in both and independent study of off-target screening for individual drugs [9] and in this study of screening chemical libraries, marketed drugs may be more target-specific and less polypharmacologic. Therefore, DrugBank may be the superior library at present for MoA-type phenotypic screening of a library.

However, annotation bias cannot be ruled out as many DrugBank libraries are annotated from individual studies and not from screening programs. Individual studies in the literature undoubtedly result in a bias toward a 'one drug, one target' datapoint, which is likely inaccurate [13]. DrugBank has the highest proportion of drugs with only one target annotated, which may explain the discrepancy between the conclusion that DrugBank is more target-specific and our results in Figure 3 which suggest the opposite. Compounds within a chemogenomic library, or a library of approved drugs are generally studied more extensively, meaning they likely have a more accurate representation of their actual pharmacology, and appear to be more polypharmacologic. Most compounds do not fall under this category, and the scarcity of this data represents the largest margin of error within our study. Excluding compounds that have no target or only one target annotated mitigates this error significantly (Table 1) but does not exclude it altogether.

The rationally designed LSP-MoA library was not significantly less polypharmacologic than the MIPE 4.0 library. Both were significantly more target-specific than the Microsource Spectrum library. All the libraries, however, showed sufficient chemical diversity to conclude that insufficient diversity was not a confounder. Calculating PP_{index} for any candidate library can alert users to the need to take polypharmacology into account, a task for which there are existing methods [8,13,15] and resources [14,16]. A large, but similar, fraction of the compounds comprising both libraries have no targets assigned at all or were not found due to incomplete data (e.g., Figure 1), Some of these are due to chiral/racemic or salt/free base variation in their database entry (Figure 4). Most targets are missing due to the lack of IC_{50}/Ki values associated with them. The targets of these drugs may have been found using

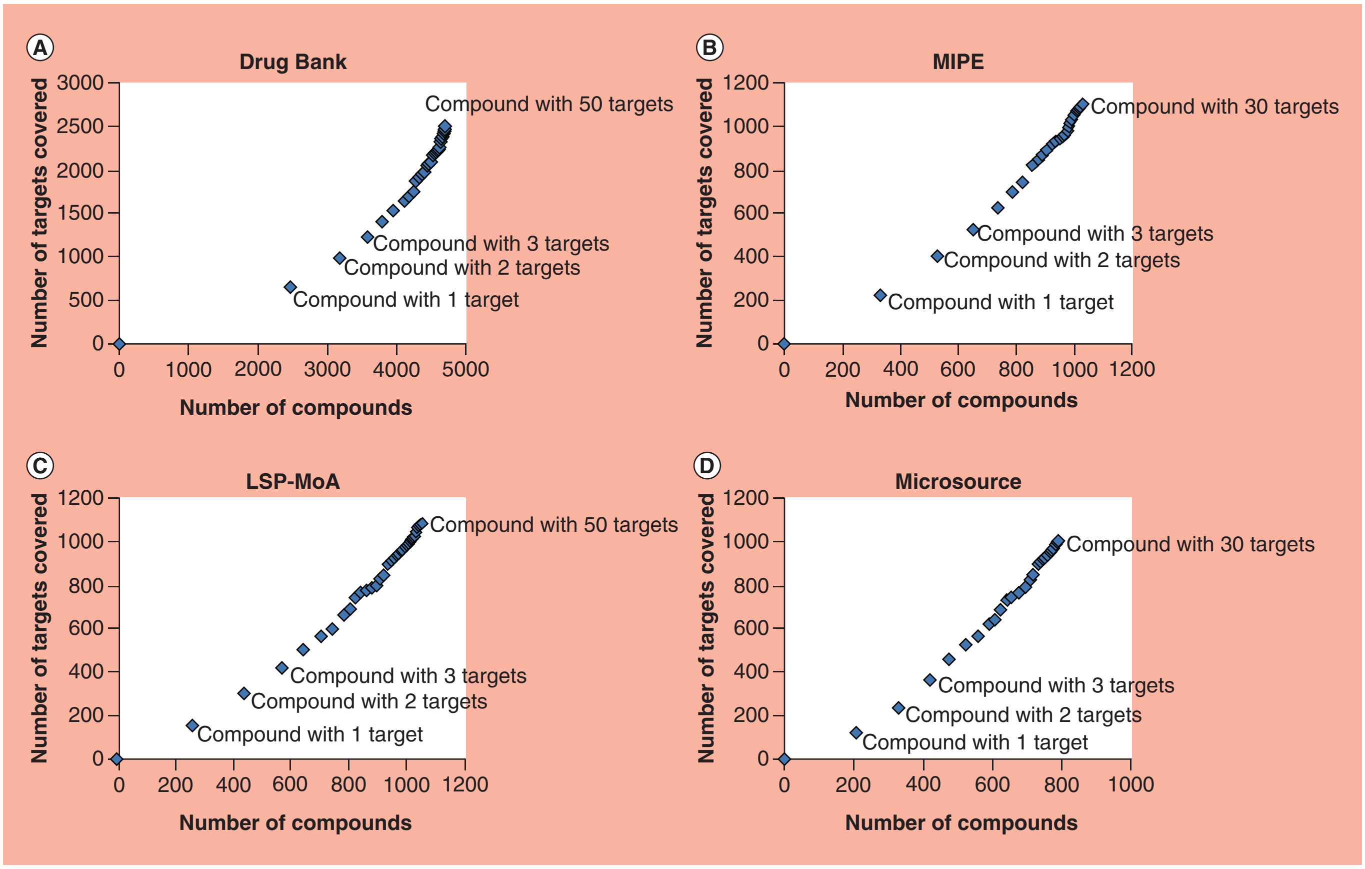


Figure 4. Confounders of chemical search. Variation in chemical form representation significantly affects search results. Examples of confounders include searches executed with stereocenters defined versus racemic entries **(A)** and molecules in a form of free base/acid against salt forms of the same molecules **(B)**.
LSP-MoA: The Laboratory of Systems Pharmacology–Method of Action; MIPE: Mechanism Interrogation PlatE.

functional/phenotypic assays or other methods for which there was no IC_{50}/Ki measurement and as such are not annotated by our method. Similarly, targets with only one compound annotated have an increased chance of bias due to many compounds being only ever tested against its one target of interest. However, these confounders represent a standard error for the search between the four libraries. PP_{index} may thus be useful in assembling and interpreting the results of phenotypic screens using chemogenomics libraries.

Future perspective

With ever more complex high-throughput biological systems such as organoid technology being developed, the reliance on phenotypic screens of small molecule drug-like compounds for drug discovery is expected to increase. The size of publicly available chemical libraries of drug-like compounds is already at the 1 billion mark and expected to increase at least linearly over time. Therefore, the reliance of successful drug discovery efforts on chemogenomics libraries is also expected to increase dramatically, as should the relationship between the probability of success of efforts using these libraries and the chemical biology quality of these libraries. Our study is an early effort to quantify the chemical biology quality of these libraries, setting the stage for the curation of more productive and efficient chemogenomics libraries in the future.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: www.future-science.com/doi/suppl/10.4144/fdd-2019-0032

Acknowledgments

We thank L Lough and O Lavinda for discussions and editorial suggestions.

Financial & competing interests disclosure

T Cardozo and K Felsovalyi are co-founders of GeneCentrix, Inc. This work was supported by NIH award R43LM012577 (to K Felsovalyi and L Young), R44DA050376 (to J Fuller) and RC2LM010994 (to T Cardozo). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Open access

Summary points

- Chemogenomics libraries assume target specificity to address the problem of target deconvolution in high-throughput phenotypic screening.
- We developed a ‘polypharmacology index’ (PP_{index}) for analysis of the target specificity of small molecule screening libraries.
- Target annotations from ChEMBL and DrugBank were compiled for several publicly available compound libraries to make frequency distributions of targets per compound.
- These frequency distributions are converted into a single number, PP_{index}, to represent the overall polypharmacology of a screening library.
- We find that chemogenomics libraries are not significantly more target-specific than general compound libraries.
- We show that these libraries can be optimized for better PP_{index}, while minimizing the target coverage loss.
- Though there is some sparsity of data for many compounds, PP_{index} can be a useful metric for analysis of phenotypic screens.

References

Papers of special note have been highlighted as: • of interest; •• of considerable interest

1. Lee J, Bogyo M. Target deconvolution techniques in modern phenotypic profiling. *Curr. Opin. Chem. Biol.* 17(1), 118–126 (2013).
2. Swinney DC, Anthony J. How were new medicines discovered? *Nat. Rev. Drug Discov.* 10(7), 507–519 (2011).

• **Meta-analysis shows that drugs discovered via phenotypic screening have a higher probability of success than those discovered via target-specific high-throughput screening.**

3. Vincent F, Loria P, Pregel M *et al.* Developing predictive assays: the phenotypic screening “rule of 3”. *Sci. Transl. Med.* 7(293), 293ps215 (2015).
4. Mathews Griner LA, Guha R, Shinn P *et al.* High-throughput combinatorial screening identifies drugs that cooperate with ibrutinib to kill activated B-cell-like diffuse large B-cell lymphoma cells. *Proc. Natl Acad. Sci. USA* 111(6), 2349–2354 (2014).
5. Patel PR, Sun H, Li SQ *et al.* Identification of potent Yes1 kinase inhibitors using a library screening approach. *Bioorg. Med. Chem. Lett.* 23(15), 4398–4403 (2013).
6. Schirle M, Jenkins JL. Identifying compound efficacy targets in phenotypic drug discovery. *Drug Discov. Today* 21(1), 82–89 (2016).
7. Jones LH, Bunnage ME. Applications of chemogenomic library screening in drug discovery. *Nat. Rev. Drug Discov.* 16(4), 285–296 (2017).

• **Commonly used chemogenomics screening libraries are reviewed in detail.**

8. Mestres J, Gregori-Puigjane E, Valverde S, Sole RV. The topology of drug-target interaction networks: implicit dependence on drug properties and target families. *Mol. Biosyst.* 5(9), 1051–1057 (2009).
9. Azzaoui K, Hamon J, Faller B *et al.* Modeling promiscuity based on *in vitro* safety pharmacology profiling data. *ChemMedChem* 2(6), 874–880 (2007).

•• **Off-target screening profiles are analyzed to assess promiscuity/polypharmacology of drugs and drug candidates, revealing that marketed drugs are less promiscuous.**

10. Moret N, Clark NA, Hafner M *et al.* Cheminformatics tools for analyzing and designing optimized small-molecule collections and libraries. *Cell Chem. Biol.* 26(5), 765.e3–777.e3 (2019).

- •• **This work undertakes a systematic effort to assess the polypharmacology and target coverage of kinase inhibitor libraries and design new libraries based on the new methods.**

11. Wishart DS, Feunang YD, Guo AC *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46(D1), D1074–D1082 (2018).
12. Mathews LA, Keller JM, Goodwin BL *et al.* A 1536-well quantitative high-throughput screen to identify compounds targeting cancer stem cells. *J. Biomol. Screen.* 17(9), 1231–1242 (2012).
13. Shmelkov E, Grigoryan A, Swetnam J *et al.* Historeceptomic fingerprints for drug-like compounds. *Front. Physiol.* 6, 371 (2015).

- • **Drug affinities are integrated with tissue expression of their targets to reveal in which tissues of the human body drugs are likely to have their greatest bioactivity.**

14. ChEMBL database (2019). www.ebi.ac.uk/chembl/
15. Gujral TS, Peshkin L, Kirschner MW. Exploiting polypharmacology for drug target deconvolution. *Proc. Natl Acad. Sci. USA* 111(13), 5048–5053 (2014).
16. GeneCentrix Inc. Historeceptomics Profiler (2019). historeceptomics.com/login