CHALLENGES IN SECONDARY ANALYSIS OF HIGH THROUGHPUT SCREENING DATA

AURORA S. BLUCHER, SHANNON K. MCWEENEY

Division of Bioinformatics and Computational Biology, Oregon Health & Science University
Portland, OR 97203 USA
Emails: blucher@ohsu.edu, mcweeney@ohsu.edu

Repurposing an existing drug for an alternative use is not only a cost effective method of development, but also a faster process due to the drug's previous clinical testing and established pharmokinetic profiles. A potentially rich resource for computational drug repositioning approaches is publically available high throughput screening data, available in databases such as PubChem Bioassay and ChemBank. We examine statistical and computational considerations for secondary analysis of publicly available high throughput screening (HTS) data with respect to metadata, data quality, and completeness. We discuss developing methods and best practices that can help to ameliorate these issues.

1. Introduction

Despite increasing investment in drug research and development in recent years, the pharmaceutical industry has seen limited results in the form of novel marketable drugs.[1] Attention has recently turned to drug repositioning, or finding new uses for already developed drugs. Drug repurposing is particularly attractive due to its simplified timeline; while the traditional drug discovery process can take between ten and seventeen years to bring a drug to production, repurposing a drug can take as little as three to twelve years depending on the drug's previously established chemical properties.[2] In several cases, repurposing has provided enormous benefit to patients with previously limited treatment options, such as the repositioning of thalidomide to treat multiple myeloma, or bromocriptine for Type 2 diabetes. Other well-known repositioning successes include Wellbutrin as Zyban for a smoking cessation aid, Minoxidil for hair loss, and Viagra (sildenafil) for erectile dysfunction.[1–3]

A potentially valuable resource for drug repositioning efforts is publically available high throughput screening (HTS) data.[4] A primary strategy for drug discovery, the automated high throughput screening process allows for the activity of hundreds of thousands of chemical compounds to be tested simultaneously.[5] Compounds are screened against a particular target compound, typically a receptor or enzyme implicated in a disease, and are declared active if their results differ from the majority of the test compounds. However, it is well known that there are several common sources of variation within high throughput screens, both technological, such as batch, plate, and positional (row or column) effects, and biological, such as the presence of non-selective binders, which can result in false positives and negative bioactivity results.[4–8] These problems are can be resolved through pre-processing, standardization and normalization methods, which include the z-score, percent inhibition, and median-based methods among others.[5,9,10]

Results from high throughput screening projects, primarily from academic institutions, are often made available through public databases such as NCBI PubChem Bioassay and ChemBank. [4] The PubChem Bioassay database contains the results of high throughput screens for the biological activities of molecules cross-listed in PubChem Substance and Compound.[11,12] Each PubChem assay has a unique assay identifier (AID). Assay data sets usually contain compound information, accompanying readout (for example, recorded fluorescence emission), activity score, activity outcome, and the mean values of minimum and maximum control wells for each plate in the assay. Activity scores and outcome are defined in the assay description, which typically explains the threshold used to declare a particular compound active.[12] The actual raw HTS data is not included in PubChem, however, and therefore there is no information on batch, plate, or within-plate position for each screened compound.

The Broad ChemBank database also contains the results of small molecule screens, as well as the raw datasets from screening centers. Each assay in ChemBank therefore contains not only compound information and accompanying readout, but also batch, plate, row, and column annotation for each screened compound. Additionally, each assay is conducted twice, so assay datasets contain replicate fluorescence readings.[13]

Given the common sources of variation known to affect high throughput screening data, it is crucial that the quality of a particular bioassay is evaluated before its results are used in further research efforts. For instance, researchers interested in using bioactivity information from databases such as PubChem and ChemBank for computational repositioning methods must first be convinced of the reliability of the screens in these databases.[7] Issues in assay quality can result in false positive or false negative bioactivity results, affecting which compounds are considered for potential repositioning. Here, datasets from both PubChem and ChemBank are evaluated to quantify the advantages and limitations of each repository as well as to investigate common sources of variation such as batch, plate, and positional effects. This analysis is representative of a typical investigation of HTS data that would be conducted before utilizing this data in further computational repurposing efforts. Overall, the problems encountered here illustrate some of the key barriers to effective secondary use of publically available high throughput screening data in order to realize the full potential of these datasets.

2. Methods

In this study, exploratory analysis was conducted on representative bioassay datasets from PubChem and ChemBank to examine data completeness, particularly in the context of data pre-processing and addressing technical sources of variation. Additional data was obtained directly from the original screeners of the highlighted PubChem study to complete the exploratory data analysis and allow for comparable assessments to the ChemBank study.

2.1 PubChem Example

The PubChem CDC25B (AID 368) dataset contains the results from approximately 65,222 compounds and controls of a primary screen against the target CDC25B. CDC25 is a protein tyrosine phosphatase cell cycle regulator, and of three existing isoforms, two are oncogenic and have been found to be overexpressed in a variety of human tumors. The goal of this screen was to find potential inhibitors for the CDC25B isoform.[14] The CDC25B dataset contained the following attributes: PubChem Substance ID, PubChem Compound ID, activity score, activity outcome, database URL, comment field, raw fluorescence intensity, calculated percent inhibition, mean of minimum control well signals (by plate), mean of maximum control well signals (by plate), calculated z-factor, and assay run date. Exploratory data analysis was conducted to evaluate the overall distribution of fluorescence intensity, percent inhibition, minimum control well means, maximum control well means, and calculated z'-factors. However, no further analysis could be performed for this dataset in the form available from the PubChem database, given the lack of plate level data such as batch number, plate number, and row and column information for each well.

2.2 Full PubChem Example

The full CDC25B dataset, including plate-level annotation, was obtained directly from the PMLSC screening center and contained results from approximately 83,711 compounds and controls across 218 384-well microtiter plates. In addition to PubChem Compound ID, raw fluorescence emission, calculated percent inhibition, mean minimum signal, mean maximum signal, calculated z-factor, and run date, this dataset also included assay batch, plate ID, row, column, well number, and well annotation. This information enabled further exploratory data analysis such as evaluation of fluorescence intensity distribution by well type and across plates and batches. Heatmaps were created for individual plates to check for positional effects. The mean signal to background ratio and percent coefficients of variation for the minimum and maximum control wells were also calculated. Based on the exploratory data analysis, percent inhibition was chosen as the most appropriate normalization method, which was also the method chosen by the original screeners when processing the dataset. [5,14]

2.3 ChemBank Example

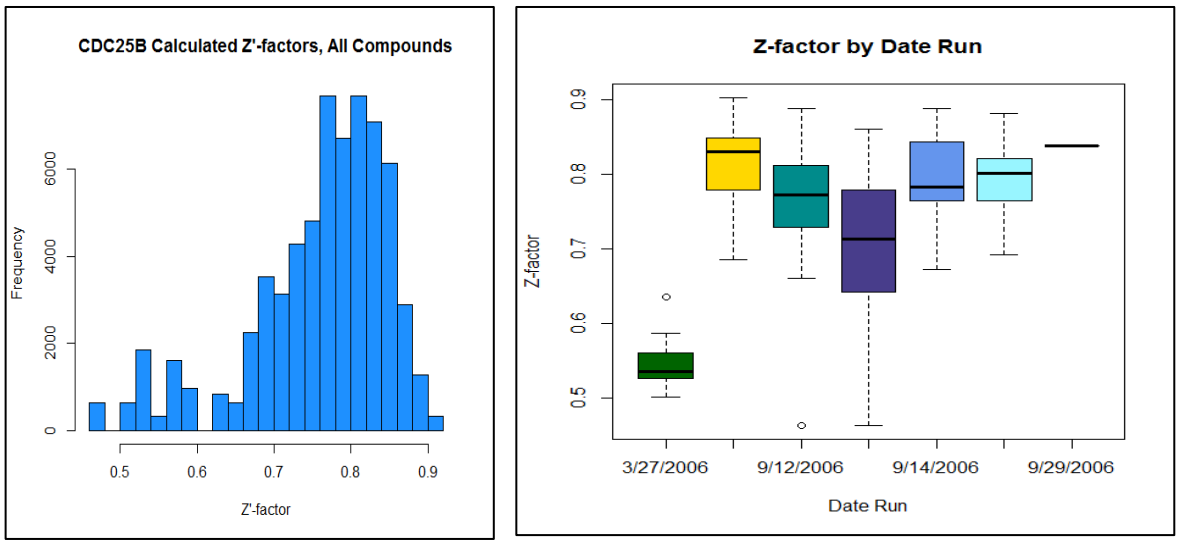
The ChemBank BRAF dataset contains the results from approximately 41,088 compounds and controls of a primary screen to find an inhibitor of the $BRAF^{V600E}$ mutant. The BRAF gene plays an important role in the mitogen-activated signaling pathway and in particular, the $BRAF^{V600E}$ mutation has been implicated in melanoma, papillary thyroid carcinoma, and colorectal cancer.[15] The BRAF dataset is composed of seven different assays, each with two replicates. Given limited assay description and annotation provided, each of the seven assays was evaluated separately. First, correlation of raw fluorescence intensity between the two replicates was assessed for each of the seven assays, and if present, any outlying data points were investigated at the plate level. Next,

exploratory data analysis was conducted for each assay to assess the overall distribution of fluorescence intensity, background-subtracted values, and calculated z-score. This analysis included histograms, boxplots, and quantile-quantile plots for individual replicates and statistical indices of the combined data, as appropriate.

3. Results

3.1 PubChem Example

Overall, the distribution of fluorescence intensity across all compounds in the CDC25B dataset is strongly skewed right, while the distribution of percent inhibition across all compounds is strongly skewed to the left. The distribution for the range between the mean minimum and mean maximum control wells is slightly skewed bimodal (See Supplementary Material S1) The distribution of z'-factors across all compounds is fairly skewed to the left and appears to be slightly bimodal. Boxplots of z'-factor by run date reveal strong variation by date (Figure 1).



(A) (B)

Figure 1. Distribution of Z'-factors for PubChem CDC25B dataset. (A) Histogram depicting distribution of calculated z'-factors. (B) Boxplots by run date for calculated z'-factors.

It is noted that the compounds run in March 2006 have much lower z'-factors than the remaining compounds, run in August and September 2006. Additionally, the compounds run on September 13th, 2006 exhibit a much wider range of z'-factors than compounds run on any other dates, while compounds run on September 29th, 2006 exhibit a much narrower range. Given that the z'-factor is a commonly used measure of assay quality, plates with a such divergent z'-factors should be examined for possible errors and batch effects. Here, however, further investigation into the sources of this variation could not be conducted due to the lack of plate level annotation available through the

PubChem Bioassay database. If the metadata had been available, it would then be possible to attempt to correct for batch and technical sources of variation.

Full PubChem CDC25B example

Histograms of fluorescence intensity by well type (compound, 50% inhibition, minimum, and maximum) for the full CDC25B dataset show that the distribution of fluorescence intensity across all wells is somewhat normal with a strong peak. The distributions of fluorescence intensities for compound wells and maximum control wells are slightly skewed right, while the distributions of fluorescence intensities for minimum and 50% inhibition control wells are more strongly skewed to the right (See Supplementary Material S2 Fig 1 and 2). Fluorescence intensity appears to vary widely by both batch and run date as well as by plate within respective batches (See Supplementary Material S2 Fig 3-8). No apparent positional effects were detected by visual examination of heatmaps for each of the 218 plates in the dataset.

Following a recently proposed decision process for HTS data processing, percent inhibition was chosen as the most appropriate method of normalization, due to the fairly normal distribution of fluorescence intensity, lack of row and column biases, a mean signal to background ratio greater than 3.5, and percent coefficients of variation for both the minimum and maximum controls wells less than 20% [5] (See Supplementary Material S2 Table 1). This appeared to successfully normalize the data by batch, date, and across plates within each batch and reproduced the original analysis (See Supplementary Material S2 Fig 9-16). It is important to note that it would not be possible to successfully evaluate this data set with regard to pre-processing and normalization without the plate level annotation.

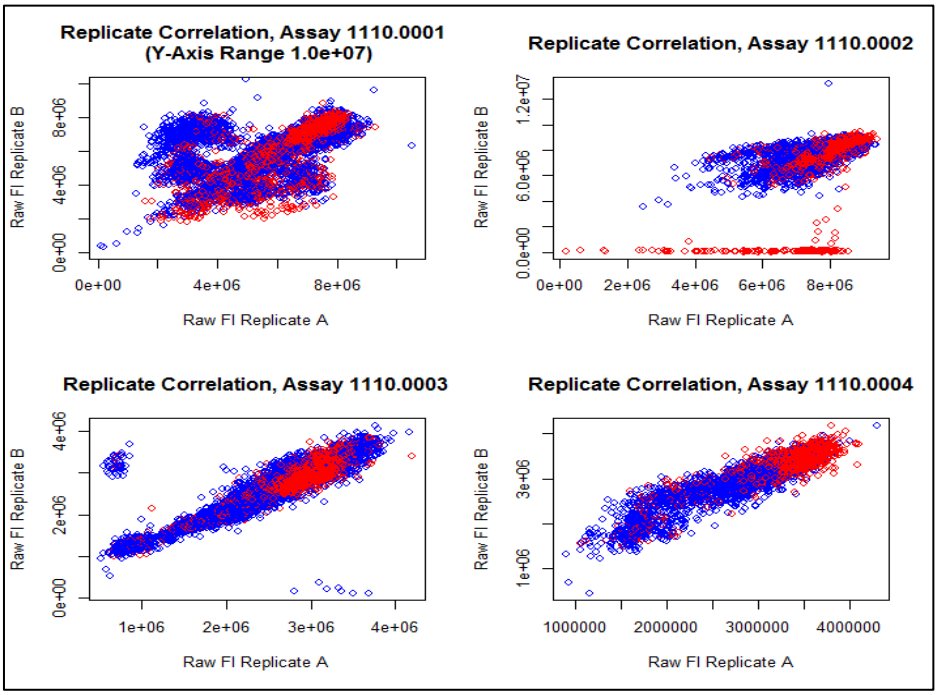
3.2 ChemBank Example

There was a large range with regard to correlation of fluorescence intensity between replicates: 0.436-0.910 (Table 1). Scatterplots further illustrate the high variability among some replicates (Figure 2). This allows easy identification of signal discrepancies. For example, the bottom of the scatterplot for assay 1110.0002, it is easy to detect a set of mock treatment wells (in red) where signal was present in replicate A, but not in replicate B. Similarly, the upper left-hand corner of the scatterplot for assay 1110.0003 shows a replicate specific cluster of compound treatment wells. The outlying data points in assay 1110.0002 were found to be confined to one plate, 1110.0002.Base. The outlying data points in assay 1110.003 were similarly located on a single plate, 1110.0003.2340.

Table 1. Correlation Coefficients for Fluorescence Intensity Replicate A vs Fluorescence Intensity Replicate B, by Assay, ChemBank BRAF dataset.

Assay Number	1110.0001	1110.0002	1110.0003	1110.0004	1110.0005	1110.0006	1110.0007
Correlation	0.436	0.536	0.906	0.910	0.902	0.869	0.846

Examination of the well-plate layout for 1110.0002 allowed identification of an obvious positional effect in the upper six rows of the plate (Figure 3). Similarly for 1110.0003, the corresponding well-plate layout illustrated a clear positional effect along the bottom two rows of the plate.



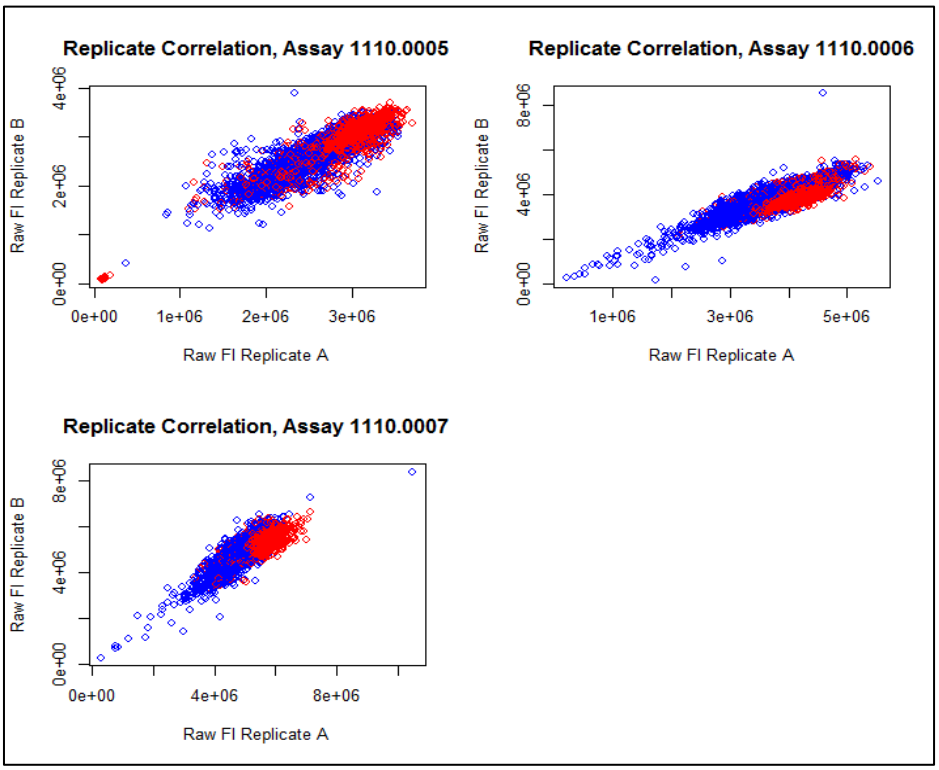


Figure 2. Scatterplots for Correlation of Fluorescence Intensity Between Replicates A and B. Correlation between replicates of Assay 1110.0001- 1110.0007. Blue indicates compound-treatment wells, red indicates control wells.

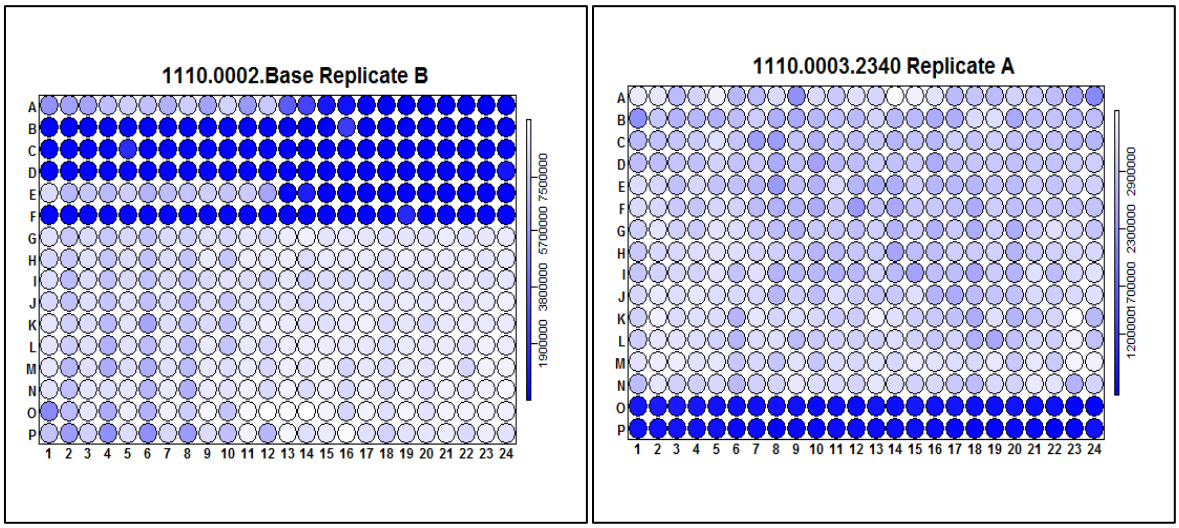


Figure 3. Well Plate Layouts for Selected BRAF Assays. (Left) Replicate B of Base Plate for Assay 1110.0003. (Right) Replicate A of Plate 2340 for Assay 1110.0003. Darker wells indicate decreased fluorescence.

Overall, each of the seven assays in the BRAF dataset showed fairly different distributions for fluorescence intensity, background-subtracted values, and calculated z-scores (See Supplementary Material S3), further reiterating the role of exploratory data analysis to examine model assumptions prior to downstream analysis.

Boxplots of the fluorescence intensity by plate were then examined. It was noted that the signal varies considerably across plates, both within and across each of the seven assays. (Replicate A shown in Figure 4). Beginning with assay 1110.0003 in replicate A, it is apparent that within each assay, fluorescence intensities steadily increase with each successive plate that is run before dropping down at the beginning of the next assay. In the absence of timestamps for each plate, it was assumed that increasing plate numbers indicate passage of time. However, without that appropriate metadata, it is not possible to determine the actual source of variation, again limiting the ability to correctly model batch or temporal effects.

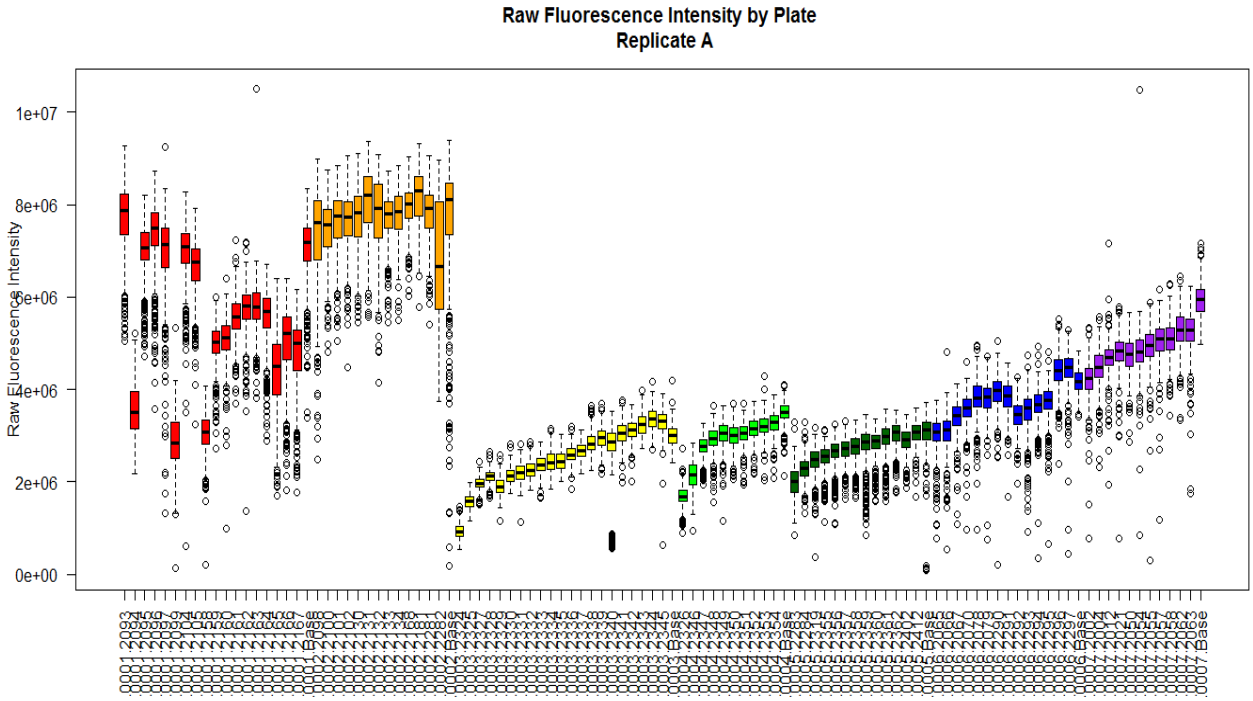


Figure 4. Raw Fluorescence Intensity by Plate, Across All Assays, Replicate A, ChemBank BRAF dataset. Each boxplot depicts the fluorescence values of the wells of one plate. Colors indicate assay "Name", which may or may not be synonymous with batch.

4. Discussion

Both repositories examined provide excellent opportunities for secondary analysis of public HTS data. However, we have noted several issues that need to be addressed in order to realize their full potential. Most notably, the lack of actual raw data, and therefore plate level annotation for bioassays in PubChem BioAssay prevents rigorous analysis of data quality. As illustrated above, initial exploratory analysis of the limited CDC25B dataset (as obtained from PubChem) reveals potential quality issues, such as variation by run date. These issues cannot be fully investigated, however, without knowledge of batch and plate numbers and row and column positioning for each tested compound. The complete CDC25B dataset, obtained directly from the screeners, allowed for more in-depth investigation of sources of variation, which in turn allowed for more appropriate pre-processing and normalization recommendations to be made. It would not have been possible to evaluate the dataset solely from the data and annotation made available through the PubChem database.

Another issue for researchers seeking to extract assay information from PubChem is the lack of description for the particular readouts used in assays. While the PubChem assay discussed in this paper provided a full description of the fluorescence emission readout, many assays do not necessarily include this level of information. It is also important to note that the issues discussed here are likely

extensible to other databases, such as ChEMBL, which contain bioactivity information from selected PubChem Bioassays.[16]

The ChemBank database is currently the only publically available bioassay database that requires the inclusion of plate level annotation in their datasets. While this information is crucial for secondary analysis, the value of the datasets in ChemBank is negatively impacted by the lack of assay annotation and description. For instance, the BRAF dataset was composed of seven different assays, but it was unclear how these differed from one another, if at all. From the assay descriptions, it appeared that only the first assay differs in its biological components, but there was no additional information as to why the remaining six assays were conducted separately. Additionally, while we might expect strong correlation between replicates for each assay, several assays exhibited exceptionally poor correlation, which casts doubt on the overall quality of the screening data. Furthermore, the lack of date or timestamps for the ChemBank data makes it impossible to confirm temporal batch effects, limiting one to data visualization by plate, with an assumption that plate order corresponds with time, as done in Figure 4.

Correspondence with PubChem confirmed that PubChem Bioassay does not require plate level annotation in uploaded datasets to the BioAssay database. It is also noted that there is no way to query for which, if any, datasets include this level of annotation (Personal communication with PubChem). ChemBank also confirmed that the "AssayName" field is used by depositors in different ways: it can be used for biologically different assays or batches of similar assays. Currently, there is no method of querying for datasets to identify those for which particular descriptive information/metadata are included (Personal Communication with ChemBank). These issues affect not only the general usability of the databases, but in particular hinder a larger-scale systematic quality analysis of HTS assays. The analysis presented here was restricted to one assay from each database primarily due to difficulties in accessibility and poor annotation.

Issues such as these in turn stymie the usage of high throughput screening data in further research efforts such as computational repositioning efforts requiring bioactivity information. There is the potential for improved data standards and development of best practices for data dissemination to improve the quality and reusability of the data in these repositories. At a minimum, the inclusion of metadata such as plate and well-level annotation will enable a more thorough secondary analysis of HTS data. Additional oversight to ensure descriptor fields for assays are completed may also encourage assay re-use. With respect to cost-benefit analysis, the potential for re-use of the data via secondary analysis far outweighs any costs due to additional data standards or metadata requirements, as the metadata has already been generated. Further impact in time/resources for depositing additional metadata can easily be mitigated by automation. One example of methods to facilitate the reporting of this metadata is a recently proposed method to first extract workflows directly from screening data in PubChem and then use the workflows to organize data within screening projects. [17]

Addressing these issues in the research community and in the requirements for submission to these repositories could improve the re-use of these data sets. A PubMed search for "PubChem" results in only 263 articles, and the more specific "PubChem BioAssay" pulls up only 51 articles. Querying for "ChemBank" returns even fewer articles, with only 17 results. For perspective, searching "GEO" brings up approximately 8480 results for Gene Expression Omnibus. While both PubChem BioAssay and ChemBank are fairly young databases and more expansive mining efforts using their datasets may still be yet to come, the annotation and data quality issues in both databases cannot be ignored as a potential barrier to dissemination. Expanded datasets as well as more rigorous quality standards are necessary to ensure the public data is truly accessible and re-usable.

5. Acknowledgements

Funding for this project was provided by the following grants: NLM (2T15LM007088-21); NIH/NCI (5P30CA069533-13, 4R00CA151457-03); NIH/NCATS (5UL1RR024140). Supplementary data is available at http:www.biodevlab.org/HTS

References

1. Dudley, J. T., Deshpande, T. & Butte, A. J. Exploiting drug-disease relationships for computational drug repositioning. *Brief. Bioinform.* **12,** 303–311 (2011).

2. Ashburn, T. T. & Thor, K. B. Drug repositioning: identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.* **3,** 673–683 (2004).

3. Pijl, H. *et al.* Bromocriptine: a novel approach to the treatment of type 2 diabetes. *Diabetes Care* **23,** 1154–1161 (2000).

4. Swamidass, S. J. Mining small-molecule screens to repurpose drugs. *Brief. Bioinform.* **12,** 327–335 (2011).

5. Shun, T. Y., Lazo, J. S., Sharlow, E. R. & Johnston, P. A. Identifying Actives from HTS Data Sets: Practical Approaches for the Selection of an Appropriate HTS Data-Processing Method and Quality Control Review. *J. Biomol. Screen.* **16,** 1–14 (2010).

6. Mayr, L. M. & Bojanic, D. Novel trends in high-throughput screening. *Curr. Opin. Pharmacol.* **9,** 580–588 (2009).

7. Xie, X.-Q. S. Exploiting PubChem for virtual screening. *Expert Opin. Drug Discov.* **5,** 1205–1220 (2010).

8. Macarron, R. *et al.* Impact of high-throughput screening in biomedical research. *Nat. Rev. Drug Discov.* **10,** 188–195 (2011).

9. Brideau, C. Improved Statistical Methods for Hit Selection in High-Throughput Screening. *J. Biomol. Screen.* **8,** 634–647 (2003).

10. Gribbon, P. Evaluating Real-Life High-Throughput Screening Data. *J. Biomol. Screen.* **10,** 99–107 (2005).

11. Li, Q., Cheng, T., Wang, Y. & Bryant, S. H. PubChem as a public resource for drug discovery. *Drug Discov. Today* **15,** 1052–1057 (2010).

12. Wang, Y. *et al.* An overview of the PubChem BioAssay resource. *Nucleic Acids Res.* **38,** D255–D266 (2009).

13. Seiler, K. P. *et al.* ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res.* **36,** D351–D359 (2007).

14. Johnston, P. A. *et al.* Cdc25B Dual-Specificity Phosphatase Inhibitors Identified in a High-Throughput Screen of the NIH Compound Library. *ASSAY Drug Dev. Technol.* **7,** 250–265 (2009).

15. Coffee, E. M. *et al.* Concomitant BRAF and PI3K/mTOR Blockade Is Required for Effective Treatment of BRAFV600E Colorectal Cancer. *Clin. Cancer Res.* **19,** 2688–2698 (2013).

16. Gaulton, A. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40,** D1100–D1107 (2011).

17. Calhoun, B. T., Browning, M. R., Chen, B. R., Bittker, J. A. & Swamidass, S. J. Automatically Detecting Workflows in PubChem. *J. Biomol. Screen.* **17,** 1071–1079 (2012).