Bioinformatics, 32(8), 2016, 1244–1246
doi: 10.1093/bioinformatics/btv723
Advance Access Publication Date: 9 December 2015
Applications Note

Gene expression

PharmacoGx: an R package for analysis of large pharmacogenomic datasets

Petr Smirnov[1,†], Zhaleh Safikhani[1,2,†], Nehme El-Hachem[3], Dong Wang[1], Adrian She[1], Catharina Olsen[1,4,5], Mark Freeman[1], Heather Selby[6,7], Deena M.A. Gendoo[1,2], Patrick Grossmann[6], Andrew H. Beck[8], Hugo J.W.L. Aerts[6], Mathieu Lupien[1,2,9], Anna Goldenberg[10,11] and Benjamin Haibe-Kains[1,2,*]

[1]Princess Margaret Cancer Centre, University Health Network, Toronto, ON, Canada, [2]Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada, [3]Institut De Recherches Cliniques De Montréal, Montreal, QC, Canada, [4]Interuniversity Institute of Bioinformatics in Brussels (IB)2, Brussels, Belgium, [5]Machine Learning Group (MLG), Department d'Informatique, Université libre de Bruxelles (ULB), Brussels, Belgium, [6]Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA, [7]Department of Bioinformatics, Boston University, Boston, MA, USA [8]Beth Israel Deaconess Medical Center, Boston, MA, USA [9]Ontario Institute of Cancer Research, Toronto, ON, Canada, [10]Hospital for Sick Children, Toronto, ON, Canada and [11]Department of Computer Science, University of Toronto, Toronto, ON, Canada

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Ziv Bar-Joseph

Received on 24 July 2015; revised on 19 November 2015; accepted on 6 December 2015

Abstract

Summary: Pharmacogenomics holds great promise for the development of biomarkers of drug response and the design of new therapeutic options, which are key challenges in precision medicine. However, such data are scattered and lack standards for efficient access and analysis, consequently preventing the realization of the full potential of pharmacogenomics. To address these issues, we implemented `PharmacoGx`, an easy-to-use, open source package for integrative analysis of multiple pharmacogenomic datasets. We demonstrate the utility of our package in comparing large drug sensitivity datasets, such as the Genomics of Drug Sensitivity in Cancer and the Cancer Cell Line Encyclopedia. Moreover, we show how to use our package to easily perform Connectivity Map analysis. With increasing availability of drug-related data, our package will open new avenues of research for meta-analysis of pharmacogenomic data.

Availability and implementation: `PharmacoGx` is implemented in R and can be easily installed on any system. The package is available from CRAN and its source code is available from GitHub.

Contact: bhaibeka@uhnresearch.ca or benjamin.haibe.kains@utoronto.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

One of the main challenges in precision medicine consists in identifying the best therapy for each patient. This is crucial in oncology where multiple cytotoxic and targeted drugs are available but their therapeutic benefits are either insufficient or limited to a subset of cancer patients. There is therefore a dire need for new anticancer drugs and robust biomarkers predictive of therapy response for individual patients. In this context, large-scale pharmacogenomic studies

could effectively achieve these goals by screening large panel of cancer cell lines using multiple drug candidates; these data are referred to as *drug sensitivity datasets*. The Genomics of Drug Sensitivity in Cancer (GDSC) (Garnett *et al.*, 2012) and the Cancer Cell Line Encyclopedia (CCLE) (Barretina *et al.*, 2012) studies have reported results of such screens, revealing several known and novel drug sensitivities and biomarkers. Subsequent evaluation, however, found only moderate inter-laboratory concordance in the drug response phenotypes, possibly due to differences in the experimental protocols used in the two studies (Haibe-Kains *et al.*, 2013; Hatzis *et al.*, 2014). Other pharmacogenomic studies, such as the Connectivity Map project (Lamb *et al.*, 2006), characterized the transcriptional changes induced by a large set of drugs; these data are referred to as *drug perturbation datasets*. For the full potential of these pharmacogenomics to be realized, new analytical approaches must be developed to best leverage the large quantity of valuable molecular and pharmacological data in the context of drug discovery and biomarker development. However, the lack of standardization of experimental protocols and annotations hinders meta-analysis of large pharmacogenomic studies.

To address these issues we developed `PharmacoGx`, an R package enabling users to download and interrogate large pharmacogenomic datasets that were extensively curated to ensure maximum overlap and consistency. `PharmacoGx` provides parallelized functions not only to assess the reproducibility of pharmacological and molecular data, but also to identify the molecular features that are consistently associated with drug effects.

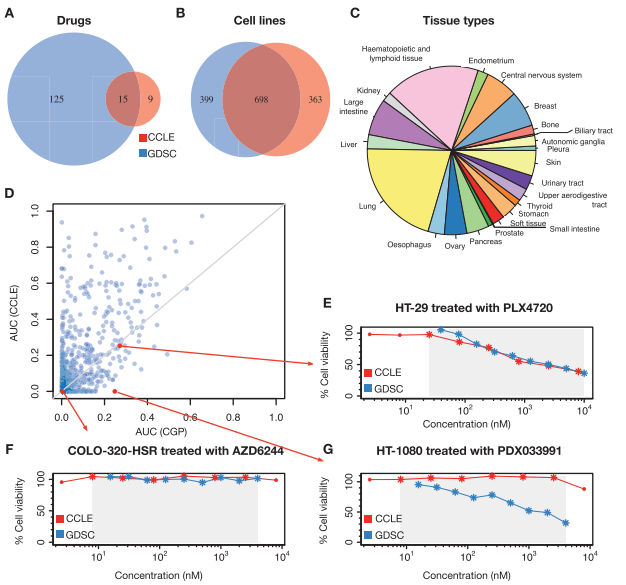


Fig. 1. Comparative study of the drug-sensitivity data across the GDSC and CCLE studies. Intersection of (**A**) drugs, (**B**) cell lines and (**C**) their tissue of origins. (**D**) Concordance of published AUC and (**E–G**) examples of concordant and discordant dose–response curves where the grey area represents the common range of concentrations

2 Package

2.1 Data

To efficiently store and analyze large pharmacogenomic datasets, we developed the `PharmacoSet` class (also referred to as `PSet`), which acts as a data container storing pharmacological and molecular data along with experimental metadata (detailed structure provided in Supplementary materials). This class enables efficient implementation of curated annotations for cell lines, drug compounds and molecular features, which facilitates comparisons between different datasets stored as `PharmacoSet` objects.

2.2 Curation of drug, cell line and molecular feature identifiers

The lack of standardization for cell line names and drug identifiers represents a major barrier for performing comparative analyses of large pharmacogenomics studies, such as GDSC and CCLE. We therefore curated these datasets to maximize the overlap in cell lines and drugs. Assigning a unique identifier to each cell line and drug, we matched entities with the same unique identifier. Manual search was then applied to match any remaining cell lines or drugs which did not match based on string similarity. Drug similarity was confirmed by examining the extended fingerprint of each of their SMILES strings and ensuring that the Tanimoto similarity between two drugs called as the same, as determined by this fingerprint, was >0.95. While standards exist for annotating genomic features, the proper mapping between microarray probe expression to genomic expression is still not entirely determined. We therefore used the BrainArray annotations, which are updated to reflect recent annotation of the human genome to perform the mapping from microarray probe to genomic expression (Dai *et al.*, 2005).

2.3 Functions

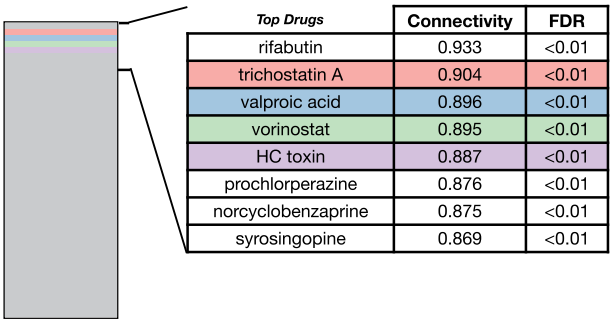
We have implemented a suite of functions facilitating the exploration and analysis of large pharmacogenomic datasets. The function `downloadPSet` allow users to directly download `PharmacoSet` objects that have been previously curated by our team. To perform comparative analysis between datasets, the lack of standards for drug and cell line identifiers must be overcome. We therefore implemented the `intersectPSets` function to make use of our extensive curation of the datasets for identifying the drugs and cell lines commonly screened in the `PharmacoSet` objects provided as input. We also included functions to explore the pharmacological measurements generated in the drug-sensitivity datasets. Full drug dose–response curves can be plotted using `drugDoseResponseCurve` and well-established drug-sensitivity estimates, such as the concentration at which the drug inhibited 50% of the maximum cellular growth (IC_{50}) and the area under the curve (AUC), can be computed using `summarizeSensitivityPhenotype`. To link molecular features to drug sensitivity the `drugSensitivitySig` function can be used to quantify the strength of each gene–drug association using a regression model controlled for treatment duration, tissue type and batch variables. Similarly the `drugPerturbationSig` function allows users to identify differential gene expressions induced by drug treatment. Finally, the `connectivityScore` function can then be used to compare drug signatures against disease signatures (tumor versus normal for instance) in order to identify drugs with carcinogenic (Caiment *et al.*, 2013) or therapeutic potential (Sirota *et al.*, 2011).

3 Case studies

We present here two case studies exploring drug sensitivity and perturbation datasets using `PharmacoGx`. The full code is provided in Supplementary material.

3.1 (In)consistency across large pharmacogenomic studies

The curated and structured aspects of our package make it easy to compare large-scale pharmacogenomic studies. We sought to



Top Drugs	Connectivity	FDR
rifabutin	0.933	<0.01
trichostatin A	0.904	<0.01
valproic acid	0.896	<0.01
vorinostat	0.895	<0.01
HC toxin	0.887	<0.01
prochlorperazine	0.876	<0.01
norcyclobenzaprine	0.875	<0.01
syrosingopine	0.869	<0.01

Fig. 2. CMAP ranking of drug connectivity scores for the HDAC inhibitor signature defined in Glaser *et al.* (2003). The four known HDAC inhibitors are ranked in the top eight drugs

reproduce and extend our published comparison of GDSC and CCLE studies (Haibe-Kains *et al.*, 2013). The most updated versions of GDSC and CCLE datasets have been curated into `PharmacoSets` that can be directly downloaded using `downloadPSet`. The 15 drugs and 698 cell lines screened in both studies can be identified using `intersectPSets` (Fig. 1A–C). Next, we can assess the concordance of all the drug-sensitivity measurements in CCLE and GDSC using the `summarizeSensitivityPhenotype` function (Fig. 1D). We can then use `drugDoseResponseCurve` to plot the (in)consistent experiments (Fig. 1E–G; Supplementary Fig. S1). In order to assess the impact of such inconsistencies in the biomarker discovery process we sought to compute the strength and significance of known gene–drug associations in the updated GDSC and CCLE datasets. We therefore used the `drugSensitivitySig` function to investigate mutations, copy number variations and gene expressions that have been reported in the literature as predictive of response to specific targeted drugs. For instance, we confirmed in GDSC and CCLE that mutation in BRAF was associated with response to the MEK inhibitor PD-0325901, that NQO1 expression was predictive of response to the HSP90 inhibitor 17-AAG and that the cell lines with ERBB2 amplification was significantly more sensitive to lapatinib (see Supplementary Material). The reproducibility of these known biomarkers supports the relevance of the GDSC and CCLE datasets despite the observed inconsistency in drug-sensitivity data.

3.2 Query the connectivity map

We further illustrated the ease of use of our package by linking drug perturbation signatures inferred from CMAP to independent signatures of HDAC inhibitors published in Glaser *et al.* (2003). We therefore sought to reproduce the HDAC analysis in Lamb *et al.* (2006) using the latest version of CMAP that can be downloaded using `downloadPSet`. The `connectivityScore` function enables the computation of the connectivity scores between the 14-gene HDAC signature from (Glaser *et al.*, 2003) and over 1000 CMAP drugs. This analysis results in the four HDAC inhibitors in CMAP being ranked at the top of the drug list (Fig. 2), therefore concurring with the original CMAP analysis (Lamb *et al.*, 2006).

4 Conclusion

The `PharmacoGx` package enables easy and efficient analysis of the increasingly available compendium of pharmacogenomic data. To the best of the authors' knowledge, this package is the first to integrate multiple pharmacogenomic datasets using structured objects incorporating standardization of cell line and drug identifiers. `PharmacoGx` includes functions to link molecular features to drug sensitivity and perturbation phenotypes, therefore providing with a unified framework to develop drug-related molecular signatures. Given that the GDSC and CCLE `PharmacoSet` objects contain multiple types of molecular profiles, which are linked to pharmacological profiles; these datasets open new avenues of research for the development of integrative biomarkers of drug response. As more datasets will be curated in `PharmacoSets`, our package will enable meta-analysis of large pharmacogenomic studies, with the aim to build better biomarkers by using multiple datasets in the discovery phase. Generation of robust biomarkers of drug response would constitute a major step toward the realization of precision medicine.

Acknowledgements

The authors would like to thank the investigators of the Genomics of Drug Sensitivity in Cancer, the Cancer Cell Line Encyclopedia and the Connectivity Map teams who have made their invaluable data available to the scientific community. The authors also thank Dr Christos Hatzis and Dr Leming Shi for their insightful discussions regarding the current obstacles in pharmacogenomics, as well as the reviewers for their constructive comments.

Funding

This work was supported by the Canadian Cancer Research Society and the Ontario Institute for Cancer Research. D.W. was supported by a CIHR-IG Computational Biology Undergraduate Summer Student Health Research Award. D.M.A.G. was supported by a CIBC-Brain Canada Brain Cancer Research Training Award. B.H.K. was supported by the Gattuso-Slaight Personalized Cancer Medicine Fund at Princess Margaret Cancer Centre.

Conflict of interest: None declared.

References

Barretina,J. *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.

Caiment,F. *et al.* (2013) Assessing compound carcinogenicity in vitro using connectivity mapping. *Carcinogenesis*, **35**, 201–207.

Dai,M. *et al.* (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.*, **33**, e175.

Garnett,M.J. *et al.* (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, **483**, 570–575.

Glaser,K.B. *et al.* (2003) Gene expression profiling of multiple histone deacetylase (HDAC) inhibitors: defining a common gene set produced by HDAC inhibition in T24 and MDA carcinoma cell lines. *Mol. Cancer Therap.*, **2**, 151–163.

Haibe-Kains,B. *et al.* (2013) Inconsistency in large pharmacogenomic studies. *Nature*, **504**, 389–393.

Hatzis,C. *et al.* (2014) Enhancing reproducibility in cancer drug screening: how do we move forward? *Cancer Res.*, **74**, 4016–4023.

Lamb,J. *et al.* (2006) The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.

Sirota,M. *et al.* (2011) Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med.*, **3**, 96ra77–96ra77.