PharmacoGx: an R package for analysis of large pharmacogenomic datasets

Petr Smirnov[1], Zhaleh Safikhani[1,2], and Benjamin Haibe-Kains[1,2]

[1]Princess Margaret Cancer Centre, University Health Network, Toronto Canada

[2]Department of Medical Biophysics, University of Toronto, Toronto Canada

July 23, 2015

Contents

1 Introduction

Pharmacogenomics hold much potential to aid in discovering drug response biomarkers and developing novel targeted therapies, leading to development of precision medicine and working towards the goal of personalized therapy. Several large experiments have been conducted, both to molecularly characterize drug dose response across many cell lines, and to examine the molecular response to drug administration. However, the experiments lack a standardization of protocols and annotations, hindering meta-analysis across several experiments.

PharmacoGx was developed to address these challenges, by providing a unified framework for downloading and analyzing large pharmacogenomic datasets which are extensively curated to ensure maximum overlap and consistency. *PharmacoGx* is based on a level of abstraction from the raw experimental data, and allows bioinformaticians and biologists to work with data at the level of genes, drugs and cell lines. This provides a more intuitive interface and, in combination with unified curation, simplifies analyses between multiple datasets.

To organize the data released by each experiment, we developed the *PharmacoSet* class. This class efficiently stores different types of data and facilitates interrogating the data by drug or cell line. The *PharmacoSet* is also versatile in its ability to deal with two distinct types of pharmacogenomic datasets. The first type, known as *sensitivity* datasets, are datasets where cell lines were profiled on the molecular level, and then tested for drug dose response. The second type of dataset is the *perturbation* dataset. These types of datasets profile a cell line on the molecular level before and after administration of a compound, to characterize the action of the compound on the molecular level.

With the first release of *PharmacoGx* we have fully curated and created PharmacoSet objects for three publicly available large pharmacogenomic datasets. Two of these datasets are of the *sensitivity* type. These are the Cancer Genome Project (CGP) [2] and the Cancer Cell Line Encyclopedia (CCLE) [1]. The third dataset is of the *perturbation* type, the Connectivity Map (CMAP) project [5].

Furthermore, PharmacoGx provides a suite of parallized functions which facilitate drug response biomarker discovery, and molecular drug characterization. This vignette will provide two example analysis case studies. The first will be comparing gene expression and drug sensitivity measures across the CCLE and CGP projects. The second case study will interrogate the CMAP database with a known signature of up and down regulated genes for HDAC inhibitors as published in [3]. Using the Connectivity Score as defined in [5], it will be seen that known HDAC inhibitors have a high numerical score and high significance.

For the purpose of this vignette, an extremely minuscule subset of all three *PharmacoSet* objects are included with the package as example data. They are included for illustrative purposes only, and the results obtained with them will likely be meaningless.

1.1 Installation and Settings

PharmacoGx requires that several packages are installed. However, all dependencies are available from CRAN or Bioconductor.

```
> source('http://bioconductor.org/biocLite.R')
> biocLite('PharmacoGx')
```

Load *PharamacoGx* into your current workspace:

```
> library(PharmacoGx)
```

Requirements

PharmacoGx has been tested on Windows, Mac and Cent OS platforms. The packages uses the core R package *parallel* to preform parallel computations, and therefore if parallelization is desired, the dependencies for the parallel package must be met.

2 Downloading PharmacoSet objects

We have made the PharmacoSet objects of the curated datasets available online at:
www.pmgenomics.ca/bhklab/software/pharmacogx/

However, to make the process of obtaining the data easier through the R shell, we have also written a function *downloadPSet* which automates downloading the datasets into a directory of the users choice, and returns the data into the R session.

```
> ## Example
> CGP <- downloadPSet("CGP")
```

Downloading Drug Signatures

The package also provides tools to compute drug perturbation and sensitivity signatures, as explained below. However, this computation is lengthy, so for users convenience we have precomputed signatures for our three PharmacoSet objects and made them available for download using the function *downloadSignatures*.

```
> ## Example
> CGP.sigs <- downloadSignatures("CGP")
```

3 Case Study

3.1 (In)Consistency across large pharmacogenomic studies

Our first case study illustrates the functions for analysis of the *sensitivity* type of dataset. The case study will investigate the consistency between the CGP and CCLE datasets, recreating the analysis similar to our *Inconsistency in Large Pharmacogenomic Studies* paper [4]. In both CCLE and CGP, the transcriptome of cells was profiled using an Affymatrix microarray chip. Cells were also tested for their response to increasing concentrations of various compounds, and form this the IC50 and AUC were computed. However, the cell and drugs names used between the two datasets were not consistent. Furthermore, two different microarray platforms were used. However, *PharmacoGx* allows us to overcome these differences to do a comparative study between these two datasets.

CGP was profiled using the hgu133a platform, while CCLE was profiled with the expanded hgu133plus2 platform. While in this case the hgu133a is almost a strict subset of hgu133plus2 platform, we can use the best information from each platform by comparing the best probes within each dataset per gene. The function *probeGeneMapping* selects a probe for each gene profiled in the microarray platform. It picks the most accurate probe for each gene using annotations extracted from the jetset package [6]. Jetset scores each probe on how sensitive and specific it is, allowing probeGeneMapping to pick the probe which is most representative of each genes true expression.

To begin, you would load the datasets from disk or download them using the *downloadPSet* function above, and then execute probeGeneMapping on the datasets.

```
>   ##Using the included example datasets
>   library(PharmacoGx)
>   data("CGPsmall")
>   data("CCLEsmall")
>   CGPsmall <- probeGeneMapping(CGPsmall)
>   CCLEsmall <- probeGeneMapping(CCLEsmall)
```

We then want to investigate the consistency of the data between the two datasets. The common intersection between the datasets can then be found using *intersectPSet*. We then create a summary of the gene expression and drug sensitivity measures for both datasets, and compare them using a standard correlation coefficient.

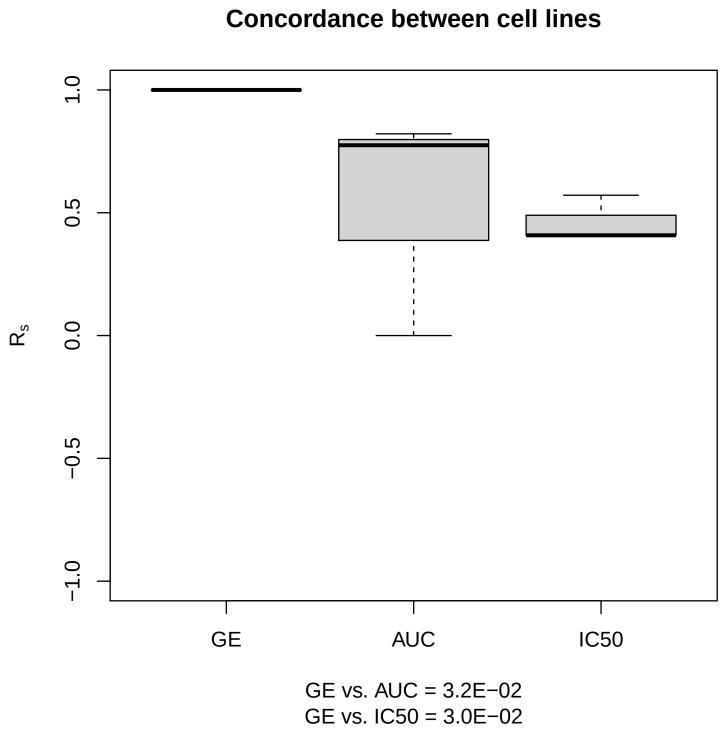
```
>   library(PharmacoGx)
>   data("CGPsmall")
>   data("CCLEsmall")
>   CGPsmall <- probeGeneMapping(CGPsmall)
>   CCLEsmall <- probeGeneMapping(CCLEsmall)
>   common <- intersectPSet(list('CCLE'=CCLEsmall,
+                                'CGP'=CGPsmall),
+                           intersectOn=c("cell.lines", "drugs", 'genes'))
>   CGP.auc <- summarizeSensitivityPhenotype(
+                 common$CGP,
+                 sensitivity.measure='auc_published',
+                 summaryStat="median")
>   CCLE.auc <- summarizeSensitivityPhenotype(
+                 common$CCLE,
+                 sensitivity.measure='auc_published',
+                 summaryStat="median")
>   CGP.ic50 <- summarizeSensitivityPhenotype(
+                 common$CGP,
+                 sensitivity.measure='ic50_published',
+                 summaryStat="median")
>   CCLE.ic50 <- summarizeSensitivityPhenotype(
+                 common$CCLE,
+                 sensitivity.measure='ic50_published',
+                 summaryStat="median")
>   common$CGP <- summarizeGeneExpression(common$CGP,
+                                         cellNames(common$CGP),
+                                         verbose=FALSE)
>   common$CCLE <- summarizeGeneExpression(common$CCLE,
+                                          cellNames(common$CCLE),
+                                          verbose=FALSE)
```

```
>   gg <- geneNames(common[[1]])
>   cc <- cellNames(common[[1]])
>   ge.cor <- sapply(cc, function (x, d1, d2) {
+     return (stats::cor(d1[ , x], d2[ , x], method="spearman",
+                 use="pairwise.complete.obs"))
+   }, d1=rnaData(common$CGP), d2=rnaData(common$CCLE))
>   ic50.cor <- sapply(cc, function (x, d1, d2) {
+     return (stats::cor(d1[, x], d2[ , x], method="spearman",
+                 use="pairwise.complete.obs"))
+   }, d1=t(CGP.ic50), d2=t(CCLE.ic50))
>   auc.cor <- sapply(cc, function (x, d1, d2) {
+     return (stats::cor(d1[ , x], d2[ , x], method="spearman",
+                 use="pairwise.complete.obs"))
+   }, d1=t(CGP.auc), d2=t(CCLE.auc))
>   w1 <- stats::wilcox.test(x=ge.cor, y=auc.cor, conf.int=TRUE, exact=FALSE)
>   w2 <- stats::wilcox.test(x=ge.cor, y=ic50.cor, conf.int=TRUE, exact=FALSE)
>   yylim <- c(-1, 1)
>   ss <- sprintf("GE vs. AUC = %.1E\nGE vs. IC50 = %.1E",
+                 w1$p.value, w2$p.value)
>   boxplot(list("GE"=ge.cor, "AUC"=auc.cor, "IC50"=ic50.cor),
+           main="Concordance between cell lines",
+           ylab=expression(R[s]),
+           sub=ss,
+           ylim=yylim,
+           col="lightgrey",
+           pch=20,
+           border="black")
>
```

3.2 Query the Connectivity Map

The second case study illustrates the analysis of a perturbation type datasets, where the changes in cellular molecular profiles are compared before and after administering a compound to the cell line. Of these datasets, we have currently curated and made available for download the Connectivity Map (CMAP) dataset [5].

For this case study, we will recreate an analysis from the paper by Lamb et al., in which a known signature for HDAC inhibitors [3] is used to recover


Concordance between cell lines
R_s
1.0
0.5
0.0
−0.5
−1.0
GE
AUC
IC50
GE vs. AUC = 3.2E−02
GE vs. IC50 = 3.0E−02

drugs in the CMAP dataset that are also known HDAC inhibitors. For this example, the package includes this signature, already mapped to the gene level, and it can be loaded by calling *data(HDAC_genes)*.

Once again, we load the dataset, and we summarize the expression on the gene level using *probeGeneMapping*. We then recreate drug signatures for each drug use the function *drugPerturbationSig* to preform statistical modelling of the transcriptomic response to the application of each drug. We then compare the observed up-regulated and down-regulated genes to a the known HDAC signature, using the GSEA connectivity score to determine the correlation between the two signatures.

```
>   library(PharmacoGx)
>   require(xtable)
>   data(CMAPsmall)
>   drug.perturbation <- drugPertubrationSig(CMAPsmall)
>   data(HDAC_genes)
>   res <- apply(drug.perturbation[,,c("tstat", "fdr")], 2, function(x, HDAC){
+             return(connectivityScore(x=x,
+                                          y=HDAC[,2,drop=FALSE],
+                                          method="gsea", nperm=100))
+         }, HDAC=HDAC_genes)
>   rownames(res) <- c("Connectivity", "P Value")
>   res <- t(res)
>   res <- res[order(res[,1], decreasing=T),]
>   xtable(res,
+     caption='Connectivity Score results for HDAC inhibitor gene signature.')
```

	Connectivity	P Value
vorinostat	0.99	0.00
alvespimycin	0.82	0.01
acetylsalicylic acid	0.50	0.26
rosiglitazone	0.00	1.00
pioglitazone	0.00	1.00

Table 1: Connectivity Score results for HDAC inhibitor gene signature.

As we can see, the known HDAC inhibitor Varinostat has a very strong connectivity score, as well as a very high significance as determined by permutation testing, in comparison to the other drugs, which score poorly.

This example serves as a validation of the method, and demonstrates the ease with which drug perturbation analysis can be done using *PharmacoGx*. While in this case we were matching a drug signature with a drug class signature, this method can also be used in the discovery of drugs that are anticorrelated with known disease signatures, to look for potential new treatments and drug repurposing.

Session Info

- R version 3.2.1 (2015-06-18), `x86_64-apple-darwin13.4.0`
- Locale: `C/en_CA.UTF-8/en_CA.UTF-8/C/en_CA.UTF-8/en_CA.UTF-8`
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Other packages: PharmacoGx 1.0.6, xtable 1.7-4
- Loaded via a namespace (and not attached): Biobase 2.28.0, BiocGenerics 0.14.0, KernSmooth 2.23-15, MASS 7.3-43, RColorBrewer 1.1-2, bitops 1.0-6, caTools 1.17.1, cluster 2.0.3, digest 0.6.8, downloader 0.4, gdata 2.17.0, gplots 2.17.0, gtools 3.5.0, igraph 1.0.1, limma 3.24.14, magicaxis 1.9.4, magrittr 1.5, marray 1.46.0, parallel 3.2.1, piano 1.8.2, plotrix 3.5-12, relations 0.6-5, sets 1.0-15, slam 0.1-32, sm 2.2-5.4, tools 3.2.1

References

[1] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A Margolin, Sungjoon Kim, Christopher J Wilson, Joseph Lehár, Gregory V Kryukov, Dmitriy Sonkin, Anupama Reddy, Manway Liu, Lauren Murray, Michael F Berger, John E Monahan, Paula Morais, Jodi Meltzer, Adam Korejwa, Judit Jané-Valbuena, Felipa A Mapa, Joseph Thibault, Eva Bric-Furlong, Pichai Raman, Aaron Shipway, Ingo H Engels, Jill Cheng, Guoying K Yu, Jianjun Yu, Peter Aspesi, Melanie de Silva, Kalpana Jagtap, Michael D Jones, Li Wang, Charles Hatton, Emanuele Palescandolo, Supriya Gupta, Scott Mahan, Carrie Sougnez, Robert C Onofrio, Ted Liefeld, Laura MacConaill, Wendy Winckler, Michael Reich, Nanxin Li, Jill P. Mesirov, Stacey B Gabriel, Gad Getz, Kristin Ardlie, Vivien Chan, Vic E Myer, Barbara L Weber,

Jeff Porter, Markus Warmuth, Peter Finan, Jennifer L Harris, Matthew Meyerson, Todd R. Golub, Michael P Morrissey, William R Sellers, Robert Schlegel, and Levi A. Garraway. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, March 2012.

[2] Mathew J Garnett, Elena J Edelman, Sonja J Heidorn, Chris D Greenman, Anahita Dastur, King Wai Lau, Patricia Greninger, I Richard Thompson, Xi Luo, Jorge Soares, Qingsong Liu, Francesco Iorio, Didier Surdez, Li Chen, Randy J Milano, Graham R Bignell, Ah T Tam, Helen Davies, Jesse A Stevenson, Syd Barthorpe, Stephen R Lutz, Fiona Kogera, Karl Lawrence, Anne McLaren-Douglas, Xeni Mitropoulos, Tatiana Mironenko, Helen Thi, Laura Richardson, Wenjun Zhou, Frances Jewitt, Tinghu Zhang, Patrick O'Brien, Jessica L Boisvert, Stacey Price, Wooyoung Hur, Wanjuan Yang, Xianming Deng, Adam Butler, Hwan Geun Choi, Jae Won Chang, Jose Baselga, Ivan Stamenkovic, Jeffrey A Engelman, Sreenath V Sharma, Olivier Delattre, Julio Saez-Rodriguez, Nathanael S Gray, Jeffrey Settleman, P Andrew Futreal, Daniel A Haber, Michael R Stratton, Sridhar Ramaswamy, Ultan McDermott, and Cyril H Benes. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391):570–575, March 2012.

[3] Keith B Glaser, Michael J Staver, Jeffrey F Waring, Joshua Stender, Roger G Ulrich, and Steven K Davidsen. 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. *Molecular cancer therapeutics*, 2(2):151–163, February 2003.

[4] Benjamin Haibe-Kains, Nehme El-Hachem, Nicolai Juul Birkbak, Andrew C Jin, Andrew H Beck, Hugo J W L Aerts, and John Quackenbush. Inconsistency in large pharmacogenomic studies. *Nature*, 504(7480):389–393, December 2013.

[5] Justin Lamb, Emily D Crawford, David Peck, Joshua W Modell, Irene C Blat, Matthew J Wrobel, Jim Lerner, Jean-Philippe Brunet, Aravind Subramanian, Kenneth N Ross, Michael Reich, Haley Hieronymus, Guo Wei, Scott A Armstrong, Stephen J Haggarty, Paul A Clemons, Ru Wei, Steven A Carr, Eric S. Lander, and Todd R. Golub. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795):1929–1935, September 2006.

[6] Qiyuan Li, Nicolai J Birkbak, Balázs Győrffy, Zoltan Szallasi, and Aron C Eklund. Jetset: selecting the optimal microarray probe set to represent a gene. *BMC Bioinformatics*, 12(1):474, 2011.