# Drug versus Disease data package

## 1 Introduction

The Drug versus Disease data (DvD) package provides reference data set for the DvD package which is a pipeline for the comparison of drug and disease gene expression profiles where negatively correlated (enriched) profiles can be used to generate hypotheses of drug-repurposing and positively correlated (enriched) profiles may be used to infer side-effects of drugs. The reference data includes disease and drug profiles, where the disease profiles were manually curated from experiments available from the Gene Expression Omnibus (GEO) and the drug profiles from Connectivity Map version 2.

### 1.1 Drug Signatures

The Connectivity Map version 2.0 rank matrix was used to generate the reference set of drug profiles for DvD. All profiles for a given compound treatment were merged as described in [**?** ], giving 1309 ranked expression profiles based on the HGU-133A platform. These profiles were then converted to gene symbols by taking the average rank where multiple probes map to the same gene and removing probes which mapped to more than one gene. These mappings were obtained from BiomaRt. Using these signatures, pairwise similarity scores were calculated using the KS running sum statistic based on the top 100 and bottom 100 genes for each profile. Affinity propagation clustering (provided by the R package apcluster) was used to create the network of drug connections. The 1309 compounds were clustered into 103 clusters, the assignments of each compound are stored in the object drugClusters. DvDdata does not contain any R code and all data objects can be accessed using the data command in R.

■results=tex■ data(drugClusters,package="DvDdata")

### 1.2 Disease Signatures

Datasets included in the DvDdata reference set contained disease versus control samples that were derived from disease-relevant primary tissues. In total, 85 disease-associated microarray experiments (disease vs control) were acquired to represent and characterise 45 distinct diseases with (3766) individual microarrays. These were obtained from NCBI GEO microarray repository http://www.ncbi.nlm.nih.gov/geo/.

The raw CEL files were normalised using rma and probes mapped to genes using the average ranking method. The pairwise similarity score matrix of these profiles was generated using the top and bottom 100 genes and Affinity propagation used to derive a network of disease connections, the classifications of each disease profile is given in the diseaseClusters data file. The 88 profiles resulted in a network of 12 clusters, the ranked profiles used to generate these networks are stored in the diseaseRL data object.

```
> #to load the disease ranked profiles
> data(diseaseRL,package="DvDdata")
```

### 1.3 Annotation

The DvD data package automatically downloads and annotates Affymetrix probe sets to HUGO gene symbols using biomaRt. The annotationlist in DvDdata gives the Affymetrix platform annotation and associated database reference in BiomaRt to allow for automatic detection and calculation.

The genes which are in the intersection of the three Affymetrix platforms supported for automatic annotation by DvD are given in the genelist object.

```
> data(annotationlist,package="DvDdata")
> #to get the HUGO genes which are included in the reference data
> data(genelist,package="DvDdata")
```

### 1.4 GEO data

The meta information is processed in DvD through the GEOquery package. This provides the experimental design, for DvD this information is used to identify explanatory factors which may be used in a regression model. The list of available factor values which are available on the GEO website http://www.ncbi.nlm.nih.gov/geo/ are stored in the GEOfactorvalues object for use by DvD.

```
> data(GEOfactorvalues,package="DvDdata")
```

### 1.5 Cytoscape Information

An associated cytoscape plug-in is available for DvD which also uses the DvDdata package. The DvDdata package contains cytodrug and cytodisease data objects which have the edges in the network along with the distance and Running sum Peak Statistic (RPS). The latter two are used as edge attributes by Cytoscape. The Running sum Peak Statistic takes values 1 or -1 where 1 indicates a positive correlation and -1 a negative correlation. The distance measure gives the strength of this correlation. This data frame is used by the DvD package to generate cytoscape sif and edge attribute files. For links out to the external web browsers, DrugBank and MeSH DvDdata also contains search compatible terms for all nodes in the reference data sets. (Note that some compounds in the connectivity map are known not to be in DrugBank).

```
> data(cytodrug,package="DvDdata")
> #to get the compound (node) names and corresponding search terms
> data(druglabels,package="DvDdata")
```

## References


[1] Hu G, Agarwal P (2009) Human Disease-Drug Network Based on Genomic Expression Profiles, *PLoS ONE*, **4(8)**: e6536.

[2] Shigemizu D, Hu Z, Hung J-H, Huang C-L, Wang Y, et al. (2012) Using Functional Signatures to Identify Repositioned Drugs for Breast, Myelogenous Leukemia and Prostate Cancer. *PLoS Comput Biol* **8(2)**: e1002347.

[3] Sirota M *et al.* (2011) Discovery and Preclinical Validation of Drug Indications Using Compendia of Public Gene Expression Data. *Sci Trans Med*, **3:96ra77**.

[4] Subramanian A *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, **102**(43), 15545-15550.

[5] Lamb J *et al.* (2006) The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science*, **313**(5795), 1929-1935.

[6] Gentleman R *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, **5**(10), R80.

[7] Parkinson et al. (2010) ArrayExpress update—an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucl. Acids Res.*,doi: 10.1093/nar/gkq1040.

[8] Edgar R, Domrachev M, Lash AE. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucl. Acids Res*, **30**(1):207-10

[R 2008] R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0

[10] Cline *et al.* (2007) Integration of biological networks and gene expression data using Cytoscape. *Nature Protocols*, **2**, 2366-2382.