**Pathways for drugs through genes**

**(i.e. drug-gene-pathway association)**

**Drug gene association**

**Resources:** Comparative Toxicogenomics Database (CTD), PharmGKB and DrugBank

Refer RelatGold/GeneDrugAssociationGoldStandard project.

**To get a list of genes associated with drugs**

*$ cut -f2 OUTPUT\_FILE\_drugsonly\_gene\_association | sort | uniq >* ***GENESLIST***

**Gene pathway association**

**Resources:** Reactome, Biocarta (or CGAP) and Gene Set Enrichment Analysis (GSEA)

**Resource: Reactome (**[**http://reactome.org/PathwayBrowser/#TOOL=AT**](http://reactome.org/PathwayBrowser/#TOOL=AT)**)**

**To retrieve pathways for genes related to human**

The list can be downloaded from **Reactome – Analysis tool (available online)**

Input is list of gene symbols (GENESLIST from Drug gene association) and output is a .csv file with gene-pathway associations. The following Linux command is used to filter the gene-pathway associations related to human.

*$ grep -i "Homo sapiens" INPUT\_FILE > OUTPUT\_FILE\_human*

**To retrieve columns with pathways and genes**

# To get the respective columns

*$ cut -f2,13 -d',' OUTPUT\_FILE\_human > OUTPUT\_FILE\_pathwayAndGenesOnly.txt*

# To replace comma delimiter with tab delimiter (between columns)

*$ sed 's/\",\"/"\t"/g' OUTPUT\_FILE\_pathwayAndGenesOnly.txt > OUTPUT\_FILE\_pathwayAndGenesOnly\_tabdelimited.txt*

# To remove the quotes surrounding the string

*$ sed 's/"//g' OUTPUT\_FILE\_pathwayAndGenesOnly\_tabdelimited.txt > OUTPUT\_FILE\_pathwayAndGenesOnly\_tabdelimited\_quotesremoved.txt*

**To rearrange the output as genes with list of pathways**

Java program GenesAndPathwaysRegrouper.java lists each gene with a list of associated pathways. The following command is used to execute the program:

*$ javac GenesAndPathwaysRegrouper.java*

*$ java GenesAndPathwaysRegrouper* *OUTPUT\_FILE\_pathwayAndGenesOnly\_tabdelimited\_quotesremoved.txt GENESLIST OUTPUT\_FILE\_REACTOME*

**To retrieve a list of genes**

All input genes may not have pathway information. The following Linux command is used to filter a list of genes that are associated with at least one pathway.

*$ grep -i -v "none" OUTPUT\_FILE\_REACTOME | cut -f1 | sort | uniq >* ***REACTOME\_GENES***

**Resource: Biocarta (**[**https://cgap.nci.nih.gov/Pathways/BioCarta\_Pathways**](https://cgap.nci.nih.gov/Pathways/BioCarta_Pathways)**)**

**To process the source file**

Hs\_GeneData.dat can be downloaded from <https://cgap.nci.nih.gov/Info/CGAPDownload>

Java program PathwaysFromCGAP.java processes the downloaded file. The following command is used to execute the program:

*$ javac PathwaysFromCGAP.java*

*$ java PathwaysFromCGAP Hs\_GeneData.dat OUTPUT\_FILE\_BIOCARTA*

The pathways/functions may not be available for all the genes. The following Linux command is used to filter the genes with at least one pathway/function.

*$ grep -i -v "none" OUTPUT\_FILE\_BIOCARTA > OUTPUT\_FILE\_BIOCARTA\_genesWithPathwaysOnly*

**To retrieve a list of genes**

*$ grep -i -v "none" OUTPUT\_FILE\_BIOCARTA\_genesWithPathwaysOnly | cut -f1 | sort | uniq >* ***BIOCARTA\_GENES***

**Combined gene-pathway output from Reactome and Biocarta**

**To retrieve list of genes**

*$ cat REATOME\_GENES BIOCARTA\_GENES | sort | uniq > GENES\_FROM\_REACTOME\_BIOCARTA*

**To retrieve the genes-pathways**

Java program GenesAndPathwayRetriever.java retrieves a list of genes and associated pathways for each gene from Reactome and Biocarta. The following command is used for executing the program:

*$ javac GenesAndPathwaysRetriever.java*

*$ java GenesAndPathwaysRetriever* *REATOME\_GENES* *BIOCARTA\_GENES* *OUTPUT\_FILE\_REACTOME* *OUTPUT\_FILE\_BIOCARTA\_genesWithPathwaysOnly* *GENES\_FROM\_REACTOME\_BIOCARTA* *GENES\_AND\_PATHWAYS\_FROM\_REACTOME\_BIOCARTA*

**To retrieve gene-pathway pairs**

Java program GenesAndPathwayPairs.java retrieves a list of gene – pathway pairs from Reactome and Biocarta. The following command is used for executing the program:

*$ javac GenesAndPathwayPairs.java*

*$ java GenesAndPathwayPairs GENES\_AND\_PATHWAYS\_FROM\_REACTOME\_BIOCARTA GENE\_PATHWAY\_PAIRS\_FROM\_REACTOME\_BIOCARTA*

**Resource: GSEA (Gene Set Enrichment Analysis) (May 2016 version)**

**C2 version - Pathways**

The pathway resource is available in c2 version and the file c2.cp.v6.0.symbols.gmt can be downloaded from <http://software.broadinstitute.org/gsea/downloads.jsp>. The pathways in GSEA are compiled from KEGG, Reactome, BioCarta and many other resources.

Java program, GenesPathwaysMiner.java processes the downloaded file.

*$ javac GenesPathwaysMiner.java*

*$ java GenesPathwaysMiner c2.cp.v6.0.symbols.gmt GSEA\_OUTPUT\_FILE\_pathways*

**C5 version – Gene ontology**

The c5 version includes gene ontology (GO). We considered c5 version to include gene-function associations. The resource file c5.all.v6.0.symbols.gmt can be downloaded from <http://software.broadinstitute.org/gsea/downloads.jsp>.

Java program GenesGOMiner.java processes the downloaded file to obtain gene and GO annotations (i.e. functions).

*$ javac GenesGOMiner.java*

*$ java GenesGOMiner c5.all.v6.0.symbols.gmt OUTPUT\_FILE*

**To get a list of gene-function association list:**

*$ cut -f2,3 OUTPUT\_FILE > GSEA\_OUTPUT\_FILE\_functions*

**C2 + C5**

C2 and C5 versions of GSEA are combined to get a unique list of gene-pathway/functions pair using the following Linux command.

*$ cat GSEA\_OUTPUT\_FILE\_pathways GSEA\_OUTPUT\_FILE\_functions | sort | uniq > GSEA\_OUTPUT\_FILE\_pathways\_and\_functions*

**Reactome + BioCarta + GSEA (c2 and c5 versions):**

**Gene-pathway/function**

*$ cat GENE\_PATHWAY\_PAIRS\_FROM\_REACTOME\_BIOCARTA GSEA\_OUTPUT\_FILE\_pathways\_and\_functions | sort -f | uniq -i > OUTPUT\_FILE\_all\_resources*

**Normalization Rule 1: Converting the case**

Java program PathwayOrFunctionToSmallLetter.java changes the pathway / function name to lower case.

*$ javac PathwayOrFunctionToSmallLetter.java*

*$ java PathwayOrFunctionToSmallLetter OUTPUT\_FILE\_all\_resources OUTPUT\_FILE*\_*lowerCase*

**Normalization Rule 2: Removing preceding and succeeding white spaces**

Java program DuplicateRecordsRemover.java trims the preceding and succeeding white spaces in gene symbol and pathway / function names.

*$ javac DuplicateRecordsRemover.java*

*$ java DuplicateRecordsRemover OUTPUT\_FILE*\_*lowerCase OUTPUT\_FILE*\_*trimWhiteSpaces*

**To include “:GO” annotation to functions**

Java program GOFunctionAnnotator.java includes “:GO” annotation to functions in gene-function association from GSEA.

*$ javac GOFunctionAnnotator.java*

*$ java GOFunctionAnnotator GSEA\_OUTPUT\_FILE\_functions OUTPUT\_FILE\_trimWhiteSpaces OUTPUT\_FILE*\_*goLabel*

**Normalization Rules for pathway names**

Java program PathwayNameNormalizer.java trims the preceding and succeeding white spaces in gene symbol and pathway / function names.

*$ javac PathwayNameNormalizer.java*

*$ java PathwayNameNormalizer OUTPUT\_FILE*\_*goLabel OUTPUT\_FILE*

**Summary (Jan 2018)**

|  |  |
| --- | --- |
| # of unique gene-pathway/function | 811,509 |
| # of genes | 17,743 |
| # of pathways | 9,190 |