**Drugs subset for drug repurposing application**

**Drugs from disease-drug association:**

**A. Drugs from US Food and Drug Administration (FDA):**

We downloaded “Drugs@FDA Download File (ZIP - 3MB)” from https://www.fda.gov/drugs/informationondrugs/ucm079750.htm. We assinged drug ID from our drugs lexicon.

$ javac DrugIDAssigner.java

$ javac DrugIDAssigner DRUGS\_LEXICON DRUGS\_FROM\_FDA OUTPUT\_FILE1

We noticed that certain FDA approved drugs are not assigned with drug ID. These are drug combinations (ex. HOMATROPINE METHYLBROMIDE; HYDROCODONE BITARTRATE and TRIPLE SULFA (SULFABENZAMIDE;SULFACETAMIDE;SULFATHIAZOLE)). In the current study we are interested in single drug repurposing, NOT drug combination repurposing OR repurposing FDA approved drugs. Therefore, we removed the drug combinations not mapping to drugs lexicon.

$ grep -v 'null' OUTPUT\_FILE1 > OUTPUT\_FILE2

We removed the duplicates with Linux command.

$ cat OUTPUT\_FILE2 | sort | uniq > OUTPUT\_FILE3

**B. Drugs from expert curated disease-drug associations (DDA):**

See our project https://github.com/CutaneousBioinf/LiteratureMiningTool/tree/master/RelatGold/Disease2DrugAssociationGoldStandard for compiling DDAs from CTD and NDF-RT.

DDAs from CTD and NDF-RT:

We devloped a Java program to assign drug ID to the drugs in DDA.

$ javac DrugIDUpdater.java

$ java DrugIDUpdater DRUGS\_LEXICON DDA\_FILE OUTPUT\_FILE4

DDAs from DrugBank:

We downloaded drugbank\_all\_full\_database.xml.zip from DrugBank and unzipped the file, full\_database.xml.

We extracting disease-drug relationship from full\_database.xml.

$ javac DiseaseDrugRelationshipExtractor.java

$ java DiseaseDrugRelationshipExtractor full\_database.xml DRUGBANK\_DRUGS

We assigned drug ID to the drugs from DrugBank.

$ javac DrugIDAssigner.java

$ java DrugIDAssigner DRUGS\_LEXICON DRUGBANK\_DRUGS OUTPUT\_FILE5

We noticed that six entities are not mapped with drug ID: NADH, ATP, MMDA, ACCLAIM, MCC, and AVAC. These are abbreviations and drugs lexicon do not include abbreviation. We removed the null records with Linux command.

$ grep -v 'null' OUTPUT\_FILE5 > OUTPUT\_FILE6

**C. Drugs from drug/biologic and gene interactions**

Resources: CTD, DrugBank and PharmGKB

See our project https://github.com/CutaneousBioinf/LiteratureMiningTool/tree/master/RelatGold/Gene2DrugAssociationGoldStandard for generating a list of expert curated drug/biologic and gene interactions.

We obtained the drugs from the expert curated drug/biologic and gene interactions.

$ cut -f1 DRUG\_OR\_BIOLOGICS\_AND\_GENE\_ASSOCIATION | sort | uniq -i > DRUG\_OR\_BIOLOGICS\_AND\_GENE\_ASSOCIATION\_DRUGSLIST

We assigned drug ID from our drugs lexicon.

$ javac DrugIDAssigner.java

$ java DrugIDAssigner DRUGS\_LEXICON DRUG\_OR\_BIOLOGICS\_AND\_GENE\_ASSOCIATION\_DRUGSLIST OUTPUT\_FILE7

**D. Drugs from chemical-genes/variants/haplotypes association**

Resource: PharmGKB

We downloaded relationships.zip from https://www.pharmgkb.org/downloads. We unzipped the files to get relationships.tsv file within relationships folder.

We got the list of chemicals/drugs associated with diseases or variants using the following Linux command.

$ cat relationships.tsv | awk -F'\t' '{print($1"\t"$2"\t"$3)}' | grep "Chemical" | sort | uniq > CHEMICALS\_LIST

We assigned drug ID to the list of chemicals/drugs obtained in the previous step.

$ javac DrugIDAssigner.java

$ java DrugIDAssigner DRUGS\_LEXICON CHEMICALS\_LIST OUTPUT\_FILE8

We observed that 25 drugs are not mapped with drug ID from our drugs lexicon. Among these, we manually updated the drug ID for 15 drugs that are synonyms of the drugs in our drugs lexicon. We identified these synonyms through google search.

‘1,7-dimethylxanthine’ was updated with ID for ‘dimethylxanthine’

‘Cotinine’ was updated with ID for ‘S-(-)-Cotinine’

‘clopidogrel active metabolite’ was updated with ID for ‘clopidogrel’.

‘aldophosphamide’ was updated with ID for ‘aldophosphamide acetal diacetate’.

‘etoposide catechol’ was updated with ID for ‘etoposide’.

‘chloroacetaldehyde’ was updated with ID for ‘trichloroacetaldehyde’.

‘tamoxifen-n-glucuronide’ was updated with ID for ‘tamoxifen’.

‘o-desmethyl-tramadol’ was updated with ID for ‘tramadol’.

‘fluoxetine glucuronide’ was updated with ID for ‘fluoxetine’.

‘desmethyl clomipramine’ was updated with ID for ‘clomipramine’.

‘deoxy-thioguanosine triphosphate’ was updated with ID for ‘2'-deoxythioguanosine’.

‘diclofenac acyl glucuronide’ was updated with ID for ‘diclofenac’.

Drugs lexicon contains ‘methadone’, NOT (S)-methadone or (R)-methadone. We updated ‘(S)-methadone’ and (R)-methadone with ‘methadone’ ID. ‘(S)-methadone’ and (R)-methadone are enantiomers (or optical isomers, a pair of molecules that are mirror images of each other)

Drugs lexicon contains ‘gemcitabine triphosphate ditriethylamine’, NOT ‘gemcitabine triphosphate’. We updated ‘gemcitabine triphosphate’ with the ID for ‘gemcitabine triphosphate ditriethylamine’.

Drugs lexicon contains ‘ivacaftor’. We updated the ID.

We omitted 10 drugs without drug ID. Among these, seven drugs were not assinged with drug ID (n-desmethyltamoxifen, 4-hydroxycyclophosphamide, difluorodeoxyuridine, corr-4a, thioguanosine triphosphate, 4-beta-hydroxycholesterol, and morphine-6-glucuronide) and three are not drugs (photodynamic therapy, radiotherapy, and hdl cholesterol).

**E. Drugs annotated in PharmGKB**

We downloaded drugs.zip from https://www.pharmgkb.org/downloads and unzipped it to get drugs.tsv file within the drugs folder.

We observed that drugs.tsv includes drugs, prodrugs, drug classes and others. I noticed that “Thyroid”, “Skeleton”, “Urine tests”, “Unclassified” “Various” are in drug classes. These don’t make sense for drug repurposing. Therefore, we filtered ONLY drugs and prodrugs from drugs.tsv.

$ cat drugs.tsv | awk -F'\t' '{if($6=="Drug" || $6=="Prodrug") print}' > DRUGS\_AND\_PRODRUGS

We filtered selected columns for further processing.

$ cut -f1,2,3,4 DRUGS\_AND\_PRODRUGS > DRUGS\_AND\_PRODRUGS\_SELECTED\_COLUMNS

We assinged drug ID to the drugs and prodrugs obtained in the previous step.

$ javac DrugIDAssigner.java

$ java DrugIDAssigner DRUGS\_LEXICON DRUGS\_AND\_PRODRUGS\_SELECTED\_COLUMNS DRUGS\_AND\_PRODRUGS\_SELECTED\_COLUMNS\_DRUG\_ID\_ADDED

We observed that 5 drugs are not assigned with drug ID. Manual curation confirmed that two of five drugs are actually drugs. The remaining three are not drugs: photodynamic therapy, radiotherapy, and vegetable oil. We manually assigned the drug ID to two drugs.

Drugs lexicon contains ‘methadone’, NOT (S)-methadone. We updated ‘(S)-methadone’ with ‘methadone’ ID.

Drugs lexicon contains ‘ivacaftor’. We updated the ID.

**F. Phenotype-drug association from CTD**

CTD includes phenotypes-drug association data. We downloaded the resource from http://ctdbase.org/downloads/#allphenotypes. File: CTD\_pheno\_term\_ixns.tsv

We filtered the associations related to human.

$ cat CTD\_pheno\_term\_ixns.tsv | awk -F'\t' '{if($8=="9606") print}' > OUTPUT\_FILE\_HUMAN

We retrieve selected columns for further processing.

$ cut -f1,2,4,5,10 OUTPUT\_FILE\_HUMAN > OUTPUT\_FILE\_HUMAN\_SELECTED\_COLUMNS

We assinged drug ID to the drugs obtained in the previous step.

$ javac DrugIDAssigner.java

$ java DrugIDAssigner DRUGS\_LEXICON OUTPUT\_FILE\_HUMAN\_SELECTED\_COLUMNS OUTPUT\_FILE9

CTD includes chemicals and drugs. We are interested only in drugs.

First, we retrieved the chemicals/drugs assigned with drug ID.

$ cat OUTPUT\_FILE9 | awk -F'\t' '{if($6!="null") print}' > OUTPUT\_FILE10

Next, we filtered the drugs based on semantic types related to antibiotics, clinical drugs, pharmacological substances and biologics.

$ grep -E '\_T121|\_T129|\_T195|\_T200' OUTPUT\_FILE10 > OUTPUT\_FILE11

**G. List of drugs associated with genes and other entities in PharmGKB**

We obtained the list from relationships.tsv file with a Java program.

We obtained a list of chemicals associated with genes

$ javac ChemicalGeneAssociationsFromPharmGKB.java

$ java ChemicalGeneAssociationsFromPharmGKB relationships.tsv relationships\_chemicalGeneAssociationsOnly.tsv

We obtained a list of chemicals associated with entities other than genes (ex. variants, haplotypes)

$ javac NotChemicalGeneAssociationsFromPharmGKB.java

$ java NotChemicalGeneAssociationsFromPharmGKB relationships.tsv relationships\_NotchemicalGeneAssociationsOnly.tsv

We noticed that some chemicals are associated with both genes and other entities. We obtained a list of chemicals associated only with genes.

$ javac ChemicalGeneONLYAssociationsFromPharmGKB.java

$ java ChemicalGeneONLYAssociationsFromPharmGKB relationships\_NotchemicalGeneAssociationsOnly.tsv relationships\_chemicalGeneAssociationsOnly.tsv relationships\_chemicalGeneONLYAssociationsOnly.tsv

We obtained a list of drugs associated with variants and haplotypes.

$ javac ChemicalVariantORHaplotypeAssociationsFromPharmGKB.java

$ java ChemicalVariantORHaplotypeAssociationsFromPharmGKB relationships.tsv relationships\_chemicalVariantORHaplotypeAssociationsOnly.tsv

We assigned drug ID to chemicals associated with variants and haplotypes. We don't consider the drugs associated with genes because, this is already processed (see C. Drugs from drug/biologic and gene interactions).

$ javac DrugIDAssigner.java

$ java DrugIDAssigner DRUGS\_LEXICON relationships\_chemicalVariantORHaplotypeAssociationsOnly.tsv OUTPUT\_FILE11

We removed the duplicate associations using Linux command.

$ cat OUTPUT\_FILE12 | sort | uniq > OUTPUT\_FILE13

We noticed that 18 chemicals are not assigned with drug ID. Among these, eight entities are not chemicals (i.e. 4-beta-hydroxycholesterol, corr-4a, difluorodeoxyuridine, hdl cholesterol, morphine-6-glucuronide, n-desmethyltamoxifen, photodynamic therapy, and radiotherapy).

We removed the chemicals without drug ID with the following Linux command.

$ grep -v 'null' OUTPUT\_FILE13 > OUTPUT\_FILE14

**Combine drugs from various resources**

$ cat OUTPUT\_FILE3 OUTPUT\_FILE6 OUTPUT\_FILE7 OUTPUT\_FILE8 DRUGS\_AND\_PRODRUGS\_SELECTED\_COLUMNS\_DRUG\_ID\_ADDED OUTPUT\_FILE11 OUTPUT\_FILE14 > OUTPUT\_FILE15

**Manual curation**

We manually removed three entities, air, water and oxygen.