Drug repurposing using gene expression signature matching

Introduction:

Drug discovery is a process spanning from the identification of disease mechanisms to the development of a drug that can be safely administered to humans. This process typically takes over a decade and only about 10-20% of the drugs succeed in being approval for market used[1]. Drug repurposing aims to side-step this process and find new indications for existing drugs. A well-known example of drug repurposing is Sildenafil, which is a PDE5 inhibitor originally developed to treat hypertension and angina, but is now widely used to treat erectile dysfunction[2]. Drug repurposing has therefore the potential to both reduce research time and cost needed to develop new drugs.

Several wet-lab and computational approaches are available to investigate drug repurposing opportunities[3]. In this practical, we aim to identify drugs candidates that can be repurposed for treating high cholesterol, by applying a signature matching approach between LDL association genes and on drug-induced gene expression signatures. The reasoning behind signature matching is the following: if a drug induces a change in gene expression inverse to the one observed in a diseased state, then this drug could potentially be used to treat the disease[4–6]. Different metrics can be used to calculate the relationship between disease state and drug signature with no one being considered as gold standard[7].

Gene expression signatures observed following drug exposure can be measured in model cell lines related to the disease of interest. This cell line model allows for the generation of large dataset containing gene expression signature stemming from the exposure of hundreds of different drugs. The Connectivity Map (CMap) is a one such drug signature database that contains the gene expression signatures of diverse drugs, profiled in a wide range of human model cell lines[8] and can be used for drug repurposing.

Disease signatures can be identified by RNA-seq experiments comparing the transcriptome between disease and healthy patients, or can be imputed from GWAS datasets using tissue-specific eQTL data, such as those from the GTEx covered extensively during the previous practical.

Another approach to create a disease signature is to infer a gene expression from genome wide association studies[4–6]. Several tools, such as MultiXcan[9] and S-PrediXcan[10], can be used to impute gene expression signatures using eQTL (expression quantitative trait loci) data. However, while QTL data can be used to infer gene expression from GWAS results, eQTL are tissue specific[11], a selection of the appropriate tissue based on prior knowledge of the disease etiology or through functional annotation of the GWAS used using tissue enrichment such as FUMA[12] is therefore necessary.

In this practical, we will perform drug repurposing by using a signature matching approach to identify drug candidates for high cholesterol. More specifically, we will impute a disease signature using the publicly available GWAS summary statistics of LDL-Cholesterol from the Global Lipids Genetics Consortium, and compare it against drug signatures from the CMap database.

Objectives:

This practical is composed of three main steps:

1. **Imputation of gene expression signatures associated with LDL cholesterol.**
   To impute a gene expression signature from a LDL cholesterol GWAS, we first need to select eQTLs from a tissue relevant to LDL cholesterol levels. We will use the FUMA tool to investigate which tissue(s) are associated with our GWAS summary and can be used for gene expression inference.

2. **Selection of genes significantly associated with LDL cholesterol.**
   In this step, we will select genes associated with LDL cholesterol to form the associated gene expression signature used for drug repurposing.

3. **Identification of drug repurposing candidates by performing signature matching.**
   To achieve this objective, our LDL cholesterol gene expression signature previously defined will be used as the query signature to conduct a signature matching analysis. We will use the precomputed drug-induced gene expression signatures from the CMap database to perform this step.

Methods used in this practical:

The overall workflow chart for the current analysis is shown below.

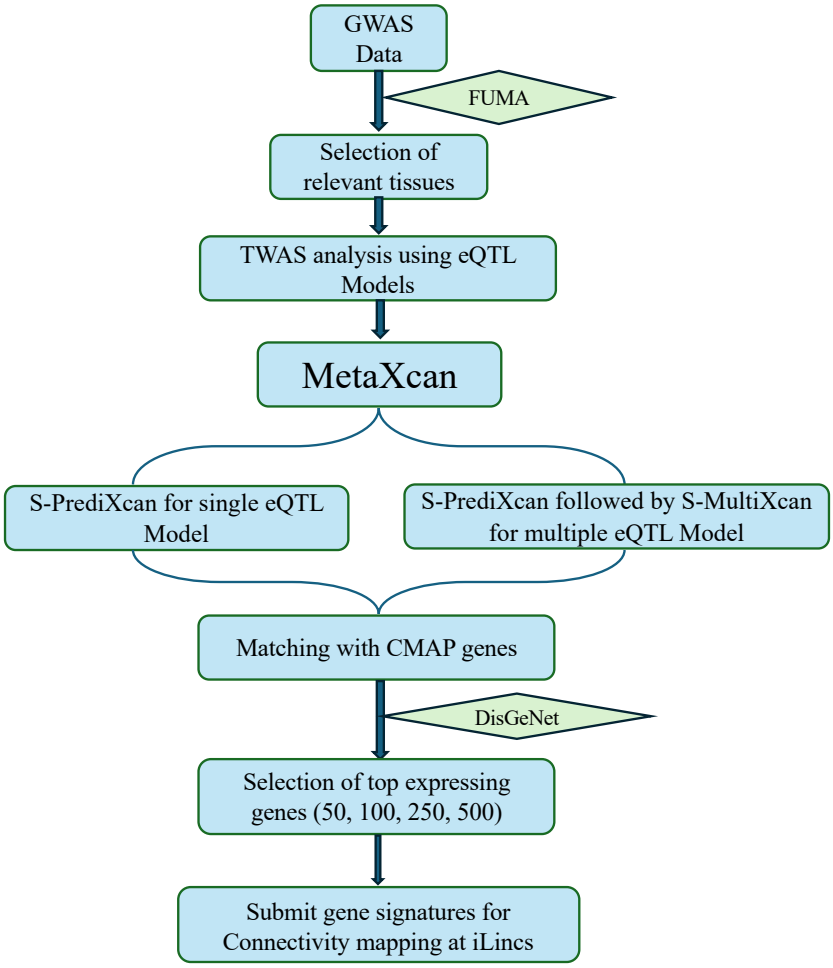


Figure 1. Step-wise workflow chart to identify drug-repurposing candidates

In this practical, we are using S-PrediXcan (a python tool) that infers a gene expression signature associated with a GWAS from eQTL stemming from a specific tissue.

While we are not performing this analysis in this practical, eQTLs from multiple tissues can be aggregated by first performing S-PrediXcan for each tissue followed by S-MultiXcan for all selected tissues (we are not using S-MultiXcan in this practical due to time limitations). This approach can be used when several tissues are associated with the phenotype of interest.

Finally, the number of genes to be included in the gene expression signatures is an arbitrary decision. In this practical, we will include 50 most upregulated and 50 most downregulated genes in our disease-associated gene expression signatures. Other choices can range from between 100 and 500 up and downregulated genes[4].

Data and tool descriptions with path for analysis:

The python tool and its anaconda environment has been installed on the cluster. The environment can be activated using the following command:

```
conda activate imlabtools
```

The data needed for this practical can be found in the following folder with the following architecture:

```
tree -d /data/module6/Practicals/Practical4_Drug_Repurposing/
```

Data analysis

1. **Tissue Selection:**

To infer the gene expression signature associated with LDL cholesterol, we first need to identify the appropriate tissue-specific eQTL model. To do this, we will utilize the FUMA (Functional Mapping and Annotation) tool that we explored in the previous practical. However, FUMA analysis is time-consuming, we have therefore already performed it and made it publicly available. The FUMA results can be accessed on the for our LDL cholesterol GWAS can be accessed on the FUMA website by following the steps below.

- Select “Browse Public Results” then type “GLGC_Willeretl” in the search tab
- Click on the “GLGC_Willeretl_Submission2” results.
- Go to the MAGMA Tissue Expression Analysis section.

Note: In case of FUMA website outage, we have saved the results locally here: /data/module6/Practicals/Practical4_Drug_Repurposing/2_Fuma_Results/LDL.jpeg

Question: Which tissue(s) is/are identified to be associated with LDL cholesterol?
Do the results make sense?

2. **Gene expression inference**:

We will infer a gene expression signature associated with LDL cholesterol using S-PrediXcan. The instructions below will get you to run this tool with the tissue of your choice.

We need to select an eQTL model to infer a gene expression signature associated with LDL cholesterol. Start by inspecting the available tissue eQTL models from GTEx using the following command:

```
ls /data/module6/Practicals/Practical4_Drug_Repurposing/3_elastic_net_models_v8/
```

Question: Based on the previous FUMA analysis and the availability of eQTL models from the GTEx database, which tissue should be selected to infer the gene expression signature associated with LDL cholesterol?

The code below will perform the S-PrediXcan analysis, modify it with the tissue of your choice as well as your output directory then run it to infer gene expression for LDL cholesterol.

```
# change the text in red to select the tissue-specific eQTL model of your choice, and to create your own output directory

conda activate imlabtools
python /software/MetaXcan/SPrediXcan.py \
--model_db_path /data/module6/Practicals/Practical4_Drug_Repurposing/3_elastic_net_models_v8/eQTLtissueModel.db \
--covariance /data/module6/Practicals/Practical4_Drug_Repurposing/3_elastic_net_models_v8/eQTLtissueModel.txt.gz \
--gwas_folder /data/module6/Practicals/Practical4_Drug_Repurposing/1_GWAS_LDL \
--gwas_file_pattern ".*txt" \
--snp_column rsID \
--effect_allele_column ALT \
--non_effect_allele_column REF \
--beta_column EFFECT_SIZE \
--pvalue_column pvalue \
--output_file /scratch/username/DirectoryToChange/LDL_model.csv

# After running the previous code, you can download the data using the following scp command
# on your local machine.
# Change the username, server IP, and directory to download the data.
scp user@203.101.225.xxx:/scratch/username/DirectoryToChange/LDL_model.csv .
```

Question: Based on the output, which genes were identified to be significantly associated with LDL cholesterol? Do the findings align with existing evidence from the literature?

3. **Defining the LDL cholesterol-associated gene expression signature:**

In this practical, we will perform a signature matching analysis using data from the CMap database. Each drug signature in the CMap database contains the gene expression changes of a total of 12,328 unique genes, induced by exposure to the specific drug of interest. This includes 978 are landmark genes (measured directly through microarray), 11,350 computationally inferred genes, of which 9,196 are inferred with high-fidelity[8]. We have retained only the 10,174 directly measured or confidently inferred genes, referred to as the Best Inferred Genes (BING), for the current analysis.

However, not all genes predicted by S-PrediXcan in the previous step have been profiled by CMap. Therefore, we need to first identify genes that are profiled in both CMap and our S-PrediXcan results and construct an LDL cholesterol signature from these genes.

First, we will have a look at the genes profiled in CMap

Run the following code in R:

```
library(corrplot)
library(dplyr)
library(reshape2)
library(ggstatsplot)
library(tidyverse)
CMAP <- read.delim(
"/data/module6/Practicals/Practical4_Drug_Repurposing/4_CMAP_Genes/GSE92742_Broad_LINCS_gene_info.txt"
) # Read the genes available in CMAP
print(nrow(CMAP)) # print the number of genes prior to filtering
CMAP <- filter(CMAP, CMAP$pr_is_bing == "1") #Select only the best inferred genes (BING)
colnames(CMAP)[2]  <- "gene_name" #Rename the gene column to be in common with the sPrediXcan data
print(nrow(CMAP)) # print the number of genes after filtering
```

Next, we will identify genes that are present in both the CMap drug signatures and the S-PrediXcan results, and rank them according to their z-score, a measurement of the association between the genes and LDL cholesterol levels.

```
Model <- read.csv("/scratch/username/DirectoryToChange/LDL_model.csv") # Read SprediXcan data
Model_cmap <- inner_join(Model, CMAP, by = "gene_name") %>% arrange(zscore)
print(nrow(Model))
print(nrow(Model_cmap))
```

Question: How many genes are profiled by both CMap and S-PrediXcan?

We will construct a signature for LDL cholesterol by selecting the top 50 most upregulated and top 50 most downregulated genes based on their association with LDL cholesterol levels (z-scores). Once constructed, this signature will be formatted to suit the requirements of iLINCs, a platform we will use to compare the LDL cholesterol signature against drug signatures.

```
Top_up50_LDL <- tail(Model_cmap, 50) #Top 50 most upregulated genes
Top_down50_LDL <- head(Model_cmap, 50) #Top 50 most downregulated genes
Gene_Signatures_LDL <- bind_rows(Top_up50_LDL, Top_down50_LDL) #Combine top expressing gene signatures
Gene_Signatures_LDL_output <- Gene_Signatures_LDL[,c(2,3,5)] # select the columns gene name, z-score and p-value
write_csv(Gene_Signatures_LDL_output
,"/scratch/username/DirectoryToChange/Gene_signatures_LDL.csv") # Write to the disk
```

Download the gene signature file from the cluster to your local system by using the following command:

```
scp user@203.101.xxx.xxx:/scratch/username/DirectoryToChange/Gene_signatures_LDL.csv .
```

4. Validation of identified signature based on disease-associated genes from DisGeNet:

DisGeNET is the largest publicly available database which consists of genes and variants associated to human diseases, collated from expert curated repositories such as GWAS catalogues, animal models[13]. We will utilize this information to validate the gene expression signature selected in the previous step. A login is required to access the DisGeNet database. Therefore, we have pre-downloaded the genes identified to be associated with hyperlipidaemia and will access them in the subsequent steps of the practical**.**

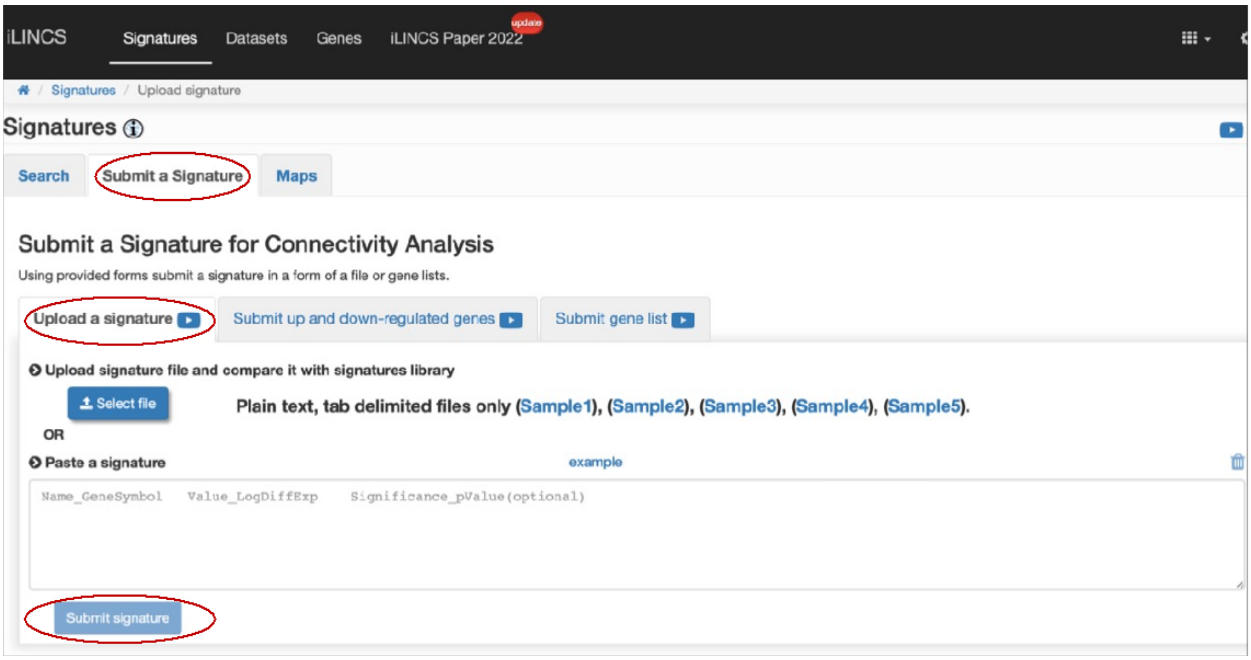
```
Hyperlipidimia_genes <-read.csv(
"/data/module6/Practicals/Practical4_Drug_Repurposing/DisGeNet_Hyperlipidimia/Hyperlipidimia_genes.csv"
) # Load genes associated with hyperlipidemia
common_genes_Model <- inner_join(Model_cmap, Hyperlipidimia_genes, by = "gene_name") %>% arrange(zscore) # identify genes in common between S-PrediXcan and DisGeNet
write.csv(common_genes_Model,
"/scratch/username/DirectoryToChange/Hyperlipidemia_Signatures_Model.csv") # write to the disk
print(dim(common_genes_Model))
common_genes_Model[1:5,]
```

Question: What genes are in common between the S-PrediXcan signature and the DisGeNet hyperlipidaemia-associated genes?

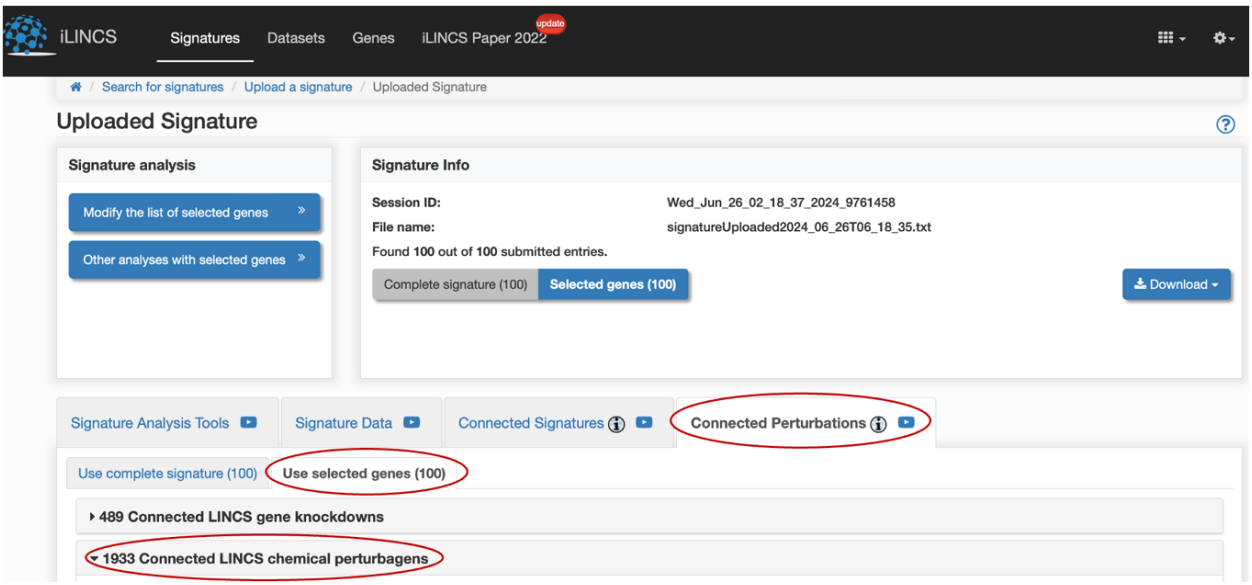
5. **Querying The CMap database with LDL cholesterol-associated gene expression signature:**

We will use the iLINCS platform to compare the LDL cholesterol signature (from our previous steps) against the drug signatures from the CMap database. iLINCS is an integrative user-friendly web platform for performing similarity analysis between user-defined gene expression signatures and the pre-computed drug perturbation signatures.

Go to the signature tab on the iLINCs website, and click on "Submit a Signature". Under "Upload signature file and compare it with signatures library", click on "select file", upload the "Gene_signatures_LDL.csv" file that we created in the previous steps, and submit signatures.



After completion of the analysis, click on "Connected Perturbations" and "Use Selected Genes". We can access the list of drugs signatures matched against our query signature by clicking on the "Connected LINCS Chemical Perturbagens".



Questions:
Are any drugs associated with our signature known to treat high cholesterol? (hint: statins)
Based on the results, what would be the most likely candidate drug for repurposing?
Is there any evidence within the literature for this drug?

Conclusions:

In this practical we went over the basic steps for drug repurposing using signature matching. We highlighted the possible use of genetic evidence by inferring gene expression signatures from GWAS. We showed that both known drugs and possible drug repurposing candidates can be identified following this approach. However, drug repurposing is often more complex especially for diseases with unknown aetiologies and requires extensive clinical validation prior to market use. Nevertheless, signature matching is a powerful tool for prioritizing drug repurposing candidates.

Extension Questions:

1. In this practical, we demonstrated the use of single tissue gene expression inference with a phenotype of known aetiology.
   How would you perform gene expression inference when disease aetiology is either unknown or involves multiple tissues?

2. Many people stop using cholesterol lowering statins due to muscle pain, a common side effect.
   How would you identify drug repurposing candidates for further investigation that could potentially provide an alternative to statins?

3. In this practical, we inferred a gene expression signature for LDL cholesterol from GWAS summary statistics.
   Can you think of alternatives that could be used to identify a disease associated gene expression signature?

References:

1. Yamaguchi, S., Kaneko, M. & Narukawa, M. Approval success rates of drug candidates based on target, action, modality, application, and their combinations. *Clin. Transl. Sci.* **14**, 1113–1122 (2021).
2. Cruz-Burgos, M. *et al.* New Approaches in Oncology for Repositioning Drugs: The Case of PDE5 Inhibitor Sildenafil. *Front. Oncol.* **11**, 627229 (2021).
3. Kulkarni, V. S., Alagarsamy, V., Solomon, V. R., Jose, P. A. & Murugesan, S. Drug Repurposing: An Effective Tool in Modern Drug Discovery. *Russ. J. bioorganic Chem.* **49**, 157–166 (2023).
4. So, H.-C. *et al.* Analysis of genome-wide association data highlights candidates for drug repositioning in psychiatry. *Nat. Neurosci.* **20**, 1342–1349 (2017).
5. Wu, P. *et al.* Integrating gene expression and clinical data to identify drug repurposing candidates for hyperlipidemia and hypertension. *Nat. Commun.* **13**, 46 (2022).
6. Reay, W. R. & Cairns, M. J. Advancing the use of genome-wide association studies for drug repurposing. *Nat. Rev. Genet.* **22**, 658–671 (2021).
7. Samart, K., Tuyishime, P., Krishnan, A. & Ravi, J. Reconciling multiple connectivity scores for drug repurposing. *Brief. Bioinform.* **22**, (2021).
8. Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **171**, 1437-1452.e17 (2017).
9. Barbeira, A. N. *et al.* Integrating predicted transcriptome from multiple tissues improves association detection. *PLoS Genet.* **15**, e1007889 (2019).
10. Barbeira, A. N. *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.* **9**, 1825 (2018).
11. Mizuno, A. & Okada, Y. Biological characterization of expression quantitative trait loci (eQTLs) showing tissue-specific opposite directional effects. *Eur. J. Hum. Genet.* **27**, 1745–1756 (2019).
12. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
13. Piñero, J. *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* **45**, D833–D839 (2017).