

1. Background

Statins, which are inhibitors of HMGCR, are cholesterol-lowering medications that are widely prescribed for treating cardiovascular diseases. There is conflicting evidence from small-scaled randomised controlled trials and observational studies on the potential anti-depressive effects of statins.

This inconsistency in findings may be due to differences in the chemical and pharmacokinetic properties of statin compounds studied, small sample sizes, varying follow-up times, unmeasured confounders and heterogeneity in depression pathophysiology. Mendelian randomisation (MR), a statistical genomic method that uses genetic instruments to proxy for drug exposure, is less prone to unmeasured confounder bias and reverse causation.

2. Aims

The aim of this practical is to use MR analysis to investigate the potential causal effects of HMGCR inhibition (the on-target effect of statins) on depression risk and symptoms. This practical is a simplified version of the analysis presented in <https://doi.org/10.1038/s41398-023-02403-8>.

More specifically, this practical is divided into three parts, with each part addressing one of the aims below:

1. identify suitable eQTLs as genetic instruments to proxy for the inhibition of HMGCR;
2. validate the genetic instruments using control traits, which are known effects of statin use;
3. investigate the association of genetically predicted HMGCR inhibition with depression risk and related traits.

3. Method Overview and Resources

We will perform the MR analysis using the SMR tool, which is available at <https://yanglab.westlake.edu.cn/software/smr/#Overview>.

We will use blood eQTLs of HMGCR, from the eQTLGen dataset (N = 31,684; available at <https://www.eqtlgen.org/phase1.html>), to proxy for HMGCR inhibition.

We will use publicly available GWAS summary statistics for diverse disease and molecular traits to investigate the association between genetically predicted HMGCR inhibition and depression-related traits.

4. Overview Of the Working Directory

Let's first have a look at our data directory. This is the folder where we store all the data necessary to perform the analysis.

```
# Let's have a look at the content of this directory  
ls /data/module6/SMR_HMGCR_prac
```

We can see that there are three sub-directories here:

Sub-directory name	Content
data	This directory contains the eQTL, LD reference and GWAS summary statistics datasets that we will use in the SMR analysis. NOTE: the README file in this directory contains the sources of each dataset, and the codes used to download and format these datasets.
output_backup	This directory contains the results of the pre-run SMR analysis. They are used as back-ups and you can compare your own results against these files if you want.
smr	This directory contains the SMR binary file that we will use in the SMR analysis.
coloc	This directory contains the R script for performing the extension analysis using coloc

Before we get started with the analysis, let's create an output directory, where we will store the MR results.

```
# Let's create the output directory
mkdir ./SMR_HMGCR_prac
cd ./SMR_HMGCR_prac
mkdir ./output/
```

5. Analysis codes

Aim 1: identification of genetic instruments

Before we do the MR analysis, we want to make sure that there are strong blood eQTLs for HMGCR. We can check this by querying (with the SMR tool) the eQTLGen dataset for any SNPs that are strongly associated with HMGCR expression.

```
# Copy and paste the following command to run an SMR query of the eQTLGen dataset
smr \
  --beqtl-summary /data/module6/SMR_HMGCR_prac/data/eQTLGen/eQTLGen_cis_eQTLs \
  --query 5.0e-8 \
  --gene ENSG00000113161 \
  --out ./output/HMGCR_eQTLs
```

While SMR is running, let's have a look at the specific flags in the command:

```
# Copy and paste the following command to run an SMR query of the eQTLGen dataset
smr \
  --beqtl-summary /data/module6/SMR_HMGCR_prac/data/eQTLGen/eQTLGen_cis_eQTLs \
  --query 5.0e-8 \
  --gene ENSG00000113161 \
  --out ./output/HMGCR_eQTLs
```

this tells the system to run the SMR tool

this tells the SMR tool to only show SNPs that are associated with HMGCR expression with a minimal p-value of 5e-8

this tells the SMR tool which eQTL dataset to use

this tells the SMR tool to only show SNPs that are associated with HMGCR (specified using its Ensembl ID)

this tells the SMR tool where to store the results

Once the analysis is completed, we can have a look at the results:

```
# We will have a look at the first ten SNPs
head ./output/HMGCR_eQTLs.txt
```

After the header, each line shows the association between one SNP and *HMGCR* expression in whole blood. Let's focus on a few columns:

Column name	Content
SNP	This is the rs ID of the SNP.
A1	This is the effect allele.
A2	This is the non-effect allele.
b	This is the effect size associated with the effect allele A1, which shows the direction and magnitude of the association between this SNP and <i>HMGCR</i> expression in whole blood.
se	This is the standard error for the effect size.
p	This is the p-value for the effect size.

As we can see, there are many SNPs that are strongly associated with *HMGCR* expression in whole blood, so we can proceed with the validation analysis.

Aim 2: Validation of genetic instruments using control traits

We need to make sure that the significant eQTLs of *HMGCR* can indeed be used to proxy for *HMGCR* inhibition. To do this, we can perform MR analysis with control traits that have previously been linked to statin exposure. As statins are cholesterol-lowering medications, we will use LDL cholesterol level as our outcome trait.

```
# Copy and paste the following command to run an SMR analysis with LDL cholesterol
smr \
  --bfile /data/module6/SMR_HMGCR_prac/data/LD_ref/1000G_phase3_20130502_combined_chr5 \
  --gwas-summary /data/module6/SMR_HMGCR_prac/data/GWAS_sumstats/Lipids_sumstats/formatted_LDL \
  --beqtl-summary /data/module6/SMR_HMGCR_prac/data/eQTLGen/eQTLGen_cis_eQTLs \
  --gene ENSG00000113161 \
  --diff-freq-prop 0.1 \
  --out ./output/HMGCR_LDL_SMR
```

While SMR is running, let's have a look at the specific flags in the command:

```
# Copy and paste the following command to run an SMR analysis with LDL cholesterol
smr \
  --bfile /data/module6/SMR_HMGCR_prac/data/LD_ref/1000G_phase3_20130502_combined_chr5 \
  --gwas-summary /data/module6/SMR_HMGCR_prac/data/GWAS_sumstats/Lipids_sumstats/formatted_LDL \
  --beqtl-summary /data/module6/SMR_HMGCR_prac/data/eQTLGen/eQTLGen_cis_eQTLs \
  --gene ENSG00000113161 \
  --diff-freq-prop 0.1 \
  --out ./output/HMGCR_LDL_SMR
```

Annotations:

- smr **: this tells the system to run the SMR tool
- bfile /data/module6/SMR_HMGCR_prac/data/LD_ref/1000G_phase3_20130502_combined_chr5 **: this tells the SMR tool which LD reference to use
- gwas-summary /data/module6/SMR_HMGCR_prac/data/GWAS_sumstats/Lipids_sumstats/formatted_LDL **: this tells the SMR tool to use the LDL summary statistic
- beqtl-summary /data/module6/SMR_HMGCR_prac/data/eQTLGen/eQTLGen_cis_eQTLs **: this tells the SMR tool to use the eQTL dataset to use
- gene ENSG00000113161 **: this tells the SMR tool which Ensembl ID
- diff-freq-prop 0.1 **: this tells the SMR tool to only perform analysis for HMGCR (specified using its Ensembl ID)
- out ./output/HMGCR_LDL_SMR**: this tells the SMR tool where to store the results
- diff-freq-prop 0.1 **: this tells the SMR tool the maximum proportion of SNPs that are allowed to have allele frequency differences

Once the analysis is completed, we can have a look at the results:

```
# Reading the result file
cat ./output/HMGCR_LDL_SMR.smr
```

Let's focus on a few columns:

Column name	Content
topSNP	This is the SNP used as the genetic instrument for the SMR analysis
A1	This is the effect allele.
A2	This is the non-effect allele.
b_SMR	This is beta, which is the estimated change in the outcome per 1 standard-deviation increase in gene expression.
se_SMR	This is the standard error for beta.
p_SMR	This is the p-value for beta.
p_HEIDI	This is the p-value for the HEIDI test.

Consider the following questions:

1. Is there a significant association between *HMGCR* expression and LDL cholesterol levels? If so, in what direction? (hint: $p_SMR < 0.05$ defines statistical significance)
2. A $p_HEIDI < 0.01$ indicates that the observed association is mediated by two or more genetic variants in LD, instead of an association mediated by one single variant. What is the p_HEIDI in this analysis? What does it mean? Does this corroborate existing evidence on the cholesterol-lowering effect of statins?

Aim 3: Analysis of the association between genetically predicted HMGCR inhibition and depression-related traits

To investigate the repurposing potential of statins on treating depression, we will repeat the SMR analysis with depression risk as an outcome.

```
# Copy and paste the following command to run an SMR analysis with depression risk
smr \
  --bfile /data/module6/SMR_HMGCR_prac/data/LD_ref/1000G_phase3_20130502_combined_chr5 \
  --gwas-summary /data/module6/SMR_HMGCR_prac/data/GWAS_sumstats/MD/formatted_MD \
  --beqtl-summary /data/module6/SMR_HMGCR_prac/data/eQTLGen/eQTLGen_cis_eQTLs \
  --gene ENSG00000113161 \
  --diff-freq-prop 0.1 \
  --out ./output/HMGCR_MD_SMR
```

Once the analysis is completed, we can have a look at the results:

```
# Reading the result file
cat ./output/HMGCR_MD_SMR.smr
```

Consider the following questions:

1. Is there a significant association between *HMGCR* expression and depression risk? If so, in what direction? (hint: $p_{\text{SMR}} < 0.05$ defines statistical significance)
2. What are some of the follow-up analyses that we can do?

Individuals with depression have been found to show increased platelet activity, with platelet activity reduced after antidepressant treatment. Interestingly, platelets are a major reservoir of serotonin in humans, and serotonin's role in depression is supported by the efficacy of selective serotonin reuptake inhibitors as a treatment of depression.

We thus repeat the SMR analysis with platelet count as an outcome.

```
# Copy and paste the following command to run an SMR analysis with depression risk
smr \
  --bfile /data/module6/SMR_HMGCR_prac/data/LD_ref/1000G_phase3_20130502_combined_chr5 \
  --gwas-summary /data/module6/SMR_HMGCR_prac/data/GWAS_sumstats/Platelet/formatted_platelet_count_nodup \
  --beqtl-summary /data/module6/SMR_HMGCR_prac/data/eQTLGen/eQTLGen_cis_eQTLs \
  --gene ENSG00000113161 \
  --diff-freq-prop 0.1 \
  --out ./output/HMGCR_PlateletCount_SMR
```

Once the analysis is completed, we can have a look at the results:

```
# Reading the result file
cat ./output/HMGCR_PlateletCount_SMR.smr
```

Consider the following questions:

1. Is there a significant association between *HMGCR* expression and platelet count? If so, in what direction? (hint: $p_{\text{SMR}} < 0.05$ defines statistical significance)
2. A $p_{\text{HEIDI}} < 0.01$ indicates that the observed association is mediated by two or more distant genetic variants in LD. What is the p_{HEIDI} in this analysis? What does it mean?

6. Extension questions

Well done on finishing the practical! You have completed your first drug repurposing analysis using SMR.

For the extension analysis, think about the following scenarios:

1. Statins have previously been reported to show off-target inhibition of ITGAL and HDAC2. Design a workflow to investigate whether these off-target effects may mediate the potential anti-depressive effects of statins.
2. By default, SMR chooses the strongest eQTL as the genetic instrument. A SNP named rs12916 has previously been used as a genetic instrument to proxy for HMGCR inhibition. Repeat the SMR analysis in the practical using rs12916 as the genetic instrument.
3. The eQTL dataset used in this practical was profiled using whole blood. Considering that the brain is biologically more relevant to the aetiology of depression, we can perform a sensitivity analysis using eQTLs generated in brain tissues. Repeat the SMR analysis in the practical using eQTL data generated in the prefrontal cortex.

(hint: replace './data/eQTLGen/eQTLGen_cis_eQTLs' with

'./data/PSYCHENCODE/PsychENCODE_cis_eqtl_HCP100_summary/Gandal_PsychENCODE_eQTL_HCP100+gPCs20_QTLtools')

4. Coloc is a tool that can be used to test whether two association datasets share a common causal variant (<https://doi.org/10.1371/journal.pgen.1004383>). As a sensitivity analysis, run the coloc analysis between *HMGCR* expression in the brain and platelet counts.

```
# Run the coloc analysis
module load r
Rscript /data/module6/SMR_HMGCR_prac/coloc/coloc.R
# Let's take a look at the results
cat ./output/HMGCR_PlateletCount_coloc_summary.csv
```

5. Pick a drug that you are interested in repurposing and design a workflow to investigate its repurposing potential to treat a disease.