Advanced Gene Mapping Course: Mendelian Randomization Exercise

Andrew DeWan, PhD, MPH

This exercise is designed to give you practical experience conducting a two-sample Mendelian randomization study using the online version of MR-base (https://www.mrbase.org/).

Part I:

You will be conducting an analysis to investigate the causal relationship between low density lipoprotein (LDL) and coronary heart disease (CHD) based on summary statistics from previously published GWAS data.

Exposure: Fasting LDL measurements from in 173,082 subjects and 2,437,752 genetic variants. Subjects are of European, East and South Asian and African ancestry.

Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, Ganna A, et al. Discovery and refinement of loci associated with lipid levels. Nat Genet. 2013 Nov;45(11):1274-1283. doi: 10.1038/ng.2797. Epub 2013 Oct 6. PMID: 24097068; PMCID: PMC3838666.

Outcome: CHD (e.g. myocardial infarction (MI), acute coronary syndrome, chronic stable angina, or coronary stenosis >50%) in 184,305 subjects (60,801 cases and 123,504 controls) and 9,455,779 genetic variants. Subjects are of European, East and South Asian, Hispanic and African ancestry.

Nikpay M, Goel A, Won HH, Hall LM, Willenborg C, Kanoni S, Saleheen D et al. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nat Genet. 2015 Oct;47(10):1121-1130. doi: 10.1038/ng.3396. Epub 2015 Sep 7. PMID: 26343387; PMCID: PMC4589895.

1) Conduct an MR analysis of LDL and CAD. Studies can be search by PubmedID in MR-base (make sure PubmedID is checked), however, please note the following:

    A. For the exposure for this publication, use the larger set of subjects for this first analysis (N=173,082)
    B. For the exposure, use a p-value threshold of 5e10-8, LD Rsq = 0.001 and clumping distance of 1000kb. Also make sure “Perform Clumping” is checked.
    C. For the outcome for this publication, use the trait denoted “Coronary heart disease”
    D. When running the MR analysis you will want to allow LD proxies to be selected for the outcome using a minimum Rsq of 0.8 and also allow for

palindromic SNPs with a MAF threshold of 0.3. Make sure you set “Allele harmonization” to “Attempt to align strands for palindromic SNPs”

E. Select the following methods:
    a. Inverse variance weighted (NOTE: this is a random effects model)
    b. MR Egger
    c. Weighted Median

Questions:

1. How many variants are included in your genetic instrument for the exposure and how many are included in the outcome analysis? Of these, how many are proxies?

2. Based on the descriptions above, is the study used to define the IV appropriate for the outcome population?

3. Is there evidence of an association between LDL and CHD?

4. Is there evidence of heterogeneity in the genetic effects?

5. Is there evidence of pleiotropy?

6. How would you interpret the results of the three analyses together (i.e. IVW, MR Egger and Weighted Median)?

2) Re-run the analysis but for myocardial infarction (MI) using outcome data from the same publication.

   Questions:

   1. Is there evidence of an association between LDL and MI?

   2. Can the association between LDL and CHD be explained by MI?

3) Feel free to explore associations with additional exposures such as HDL, BMI (you can use the Yengo et al. SNPs) or other exposures/outcomes of interest to you.

Part II:

Let's now see if we can validate the finding of an association between LDL and CHD by using different exposure data source and potentially dissect this signal to see if we pinpoint the features of LDL that might be driving this signal. We will use metabolomics data that was generated in a sample of 24,925 individuals.

Kettunen, J., Demirkan, A., Würtz, P. *et al.* Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of *LPA*. *Nat Commun* **7,** 11122 (2016). https://doi.org/10.1038/ncomms11122.

Exposures: LDL.C, LDL.D, S.LDL.C, S.LDL.L, S.LDL.P, M.LDL.C, M.LDL.CE, M.LDL.L, M.LDL.P, M.LDL.PL, L.LDL.C, L.LDL.CE, L.LDL.FC, L.LDL.L, L.LDL.P, L.LDL.PL (16 metabolites)

Where S. = small, M. = medium and L. = large; .C = total cholesterol, .D = diameter, .L = total lipids, .P = concentration, .CE = cholesterol esters, .PL = phospholipids

These can all be selected when you are on the "Choose Exposure" screen and selecting the "Metabolite level QTLs". You can then type in "LDL" in the analyte window and select each of these in the window that pops up. The most efficient way is to select all of the metabolites you're interested in and then run the MR analyses together. Before clicking off of this screen you will want to click on the "Select All" under Row Selection. This will allow you to run the analysis on all of the SNPs for each metabolite.

Use the same CHD outcome as you did for Part I (Nikpay PMID: 26343387) using the full set of cases and controls (N=184,305).

1) Conduct an MR analysis as you did previously.

   Questions:

   1. For LDL.C, does the association between LDL and CHD validate the previous findings?

   2. Considering all of the associations, do these results differentiate between the different characteristics of LDL (Please note: you will want to take into account the 16 association tests)?

3. What might be one explanation for the similarity between results for the different LDL characteristics?

4. Are there any concerns about heterogeneity or pleiotropy?

Let’s now look at the associations with VLDL metabolites using the same exposure and outcome data sources.

Exposures: 33 VLDL metabolites (Please note additional abbreviations: XS. = very small, XL. = very large, XXL. = extremely large; .TG = triglycerides)

2) Conduct an MR analysis as you did previously.

   Question:

   5. Considering all of the associations, are there any obvious trends in the results (Please note: you will want to take into account the 33 association tests)?