**Supplementary documents**

**MR-TRYX: Exploiting horizontal pleiotropy to infer novel causal pathways**

Yoonsu Cho1, Philip C Haycock1, Tom R Gaunt1, Jie Zheng1, Andrew P Morris2, George Davey Smith1, Gibran Hemani1

1 MRC Integrative Epidemiology Unit, Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK.

2 Department of Biostatistics, University of Liverpool, Liverpool, UK.

**Supplementary simulations**

In MR-TRYX, we applied a LASSO to automate the removal of redundant traits in the MR-BASE database. To validate the utility of the LASSO for removing redundant traits, we performed further simulations analyses.

**Supplementary simulation 1. The validity of LASSO step in MR-TRYX**

In this simulation, we generated data on 100 genetic variants, candidate traits X1 and X2, and outcome Y for 100,000 individuals in a two-sample MR context (Supplementary Figure S1). Each variant was bi-allelic with minor allele frequency of 0.5. The genetic effects on the trait X1 and X2 were generated from a normal distribution with effect size of 0.2, respectively. For each trait, we created 5 proxy traits that are similar to the original traits (e.g. X1a, X1b, X1c, X1d, X1e and X2a, X2b, X2c, X2d, X2e). Two sets of scenarios are considered: in the first, all instruments are valid (no pleiotropy); and in the second, we allowed some variants to be pleiotropic. In each scenario, we performed normal MVMR and ALSSO MVMR. Thousand simulated datasets were generated for each scenario.

**Supplementary simulation 2. The number of instruments and selection of the traits by LASSO**

In this simulation, we generated traits X1 and X2 which are instrumented by 100 genetic variants and 20 genetic variants, respectively. Each variant was bi-allelic with minor allele frequency of 0.5. Among 120 variants, 10 of them were considered as pleiotropic SNPs, having associations with X1 as well as X2. We considered four scenarios.

1. X1 has an effect on Y (β=0.3), where X2 has no effect on Y (β=0.0)
2. X1 has no effect on Y (β=0.0), where X2 has an effect on Y (β=0.3)
3. Both of X1 and X2 have effects on Y (β=0.3)
4. Neither of X1 and X2 have effects on Y (β=0.0)

In each scenario, we performed the LASSO regression to evaluate if the LASSO selects the traits in a manner that is determined by which trait is causal, rather than which trait has more instruments. The simulation was replicated 1000 times in each scenario.

**Supplementary results**

**Supplementary simulation 1.**

The results showed that the LASSO successfully removed redundant traits. Among 6 traits (1 original trait and 5 proxies), LASSO selected two traits on average (Supplementary Figure S2). However, we found that the LASSO does not always select the original traits X1 and X2. For example, sometimes LASSO selected proxies (e.g. X1b and X1c) but not the original trait (X1). Although the original trait is not always chosen by LASSO, our adjustment method through LASSO-MVMR is still valid as LASSO captured X1 through its proxies, X1b and X1c. Suppose X1b and X1c influences the outcome Y with effect sizes of 0.164 and 0.036, respectively. Our adjustment method will correct for both X1b and X1c (0.164+0.036 = 0.200), which is the same as the effect of X1 on Y. We demonstrated whether the sum of selected traits is the same as the effect of targeted traits (e.g. candidate traits may cause biased estimation) on the outcome through the simulation. The supplementary table S3 presents that the sum of effect size of selected trait for X1s was 0.195 (95% confidence interval, CI: 0.191, 0.198), which is almost same as given effect size, 0.200. Whilst, the effect of the X1s on the outcome was 0.420 (95% CI: 0.389, 0.450) in normal MVMR model. The sum of effects of selected traits for X2 in LASSO-MVMR was 0.192 (95% CI: 0.189, -.195), where the effect of the X2s was 0.019 (95% CI: -0,050, 0.012) in normal MVMR. This result was replicated in scenario 2, where horizontal pleiotropy exists. This simulation confirms that the LASSO MVMR corrects for effect of the candidate traits appropriately. As MR-TRYX repeats this step 1000 times, the LASSO MVMR in MR-TRYX would yield more precise estimates than MVMR on its own.

**Supplementary simulation 2.**

The results showed that the LASSO step worked well regardless of the number of instruments for the traits (Supplementary Table S4). The LASSO kept both X1 and X2 when both traits (Scenario 3) have effects on the outcome Y with a probability of 1.000 (1000 times per 1000 simulations). Whilst, the LASSO removed both traits with a probability of 0.693 when both of X1 and X2 have no effects on Y (Scenario 4). In this scenario, the trait X1 was removed with a probability of 0.775 where the trait X2 was removed with a probability of 0.783. In the scenario 1, LASSO kept the trait X1 with a 100% of probability. However, LASSO failed sometime to remove X2, which is instrumented by a smaller number of variants, with a probability of 0.473. In scenario 2, X2 wasn’t removed by LASSO step but X1 was remained with a probability of 0.467. Considering the similar probability of the trait being eliminated in each scenario, it can be suggested that the LASSO may not favour to the candidate traits with a larger number of instruments, but it removes a trait weakly associated with the outcome.

-----

heterogeneity: For every SNP, adjust for each potential p independently and keep just the best one based on the biggest reduction in heterogeneity. none: For every SNP, adjust for each potential p independently and keep Now keep all of them. mv: For each SNP, adjust for all potential p jointly by performing multivariable MR to obtain p->y causal estimates. mv\_lasso: For each SNP, adjust for all potential p jointly by selecting p based on lasso, and then using multivariable MR on remaining variables to obtain p->y causal estimates. lasso\_selection: For each SNP, adjust for all potential p jointly by selecting p based on lasso and using the original estimates. lasso: For each SNP, adjust for all potential p jointly by selecting p based on lasso and using the lasso p->y estimates with the original standard errors

1. For every SNP, adjust for each potential p independently

a. Now keep just the best one based on the biggest reduction in heterogeneity

b. Now keep all of them

2. For each SNP, adjust for all potential p jointly

a. By performing multivariable MR to obtain p->y causal estimates

b. By selecting p based on lasso, and then using multivariable MR on remaining variables to obtain p->y causal estimates

c. By selecting p based on lasso and using the original estimates

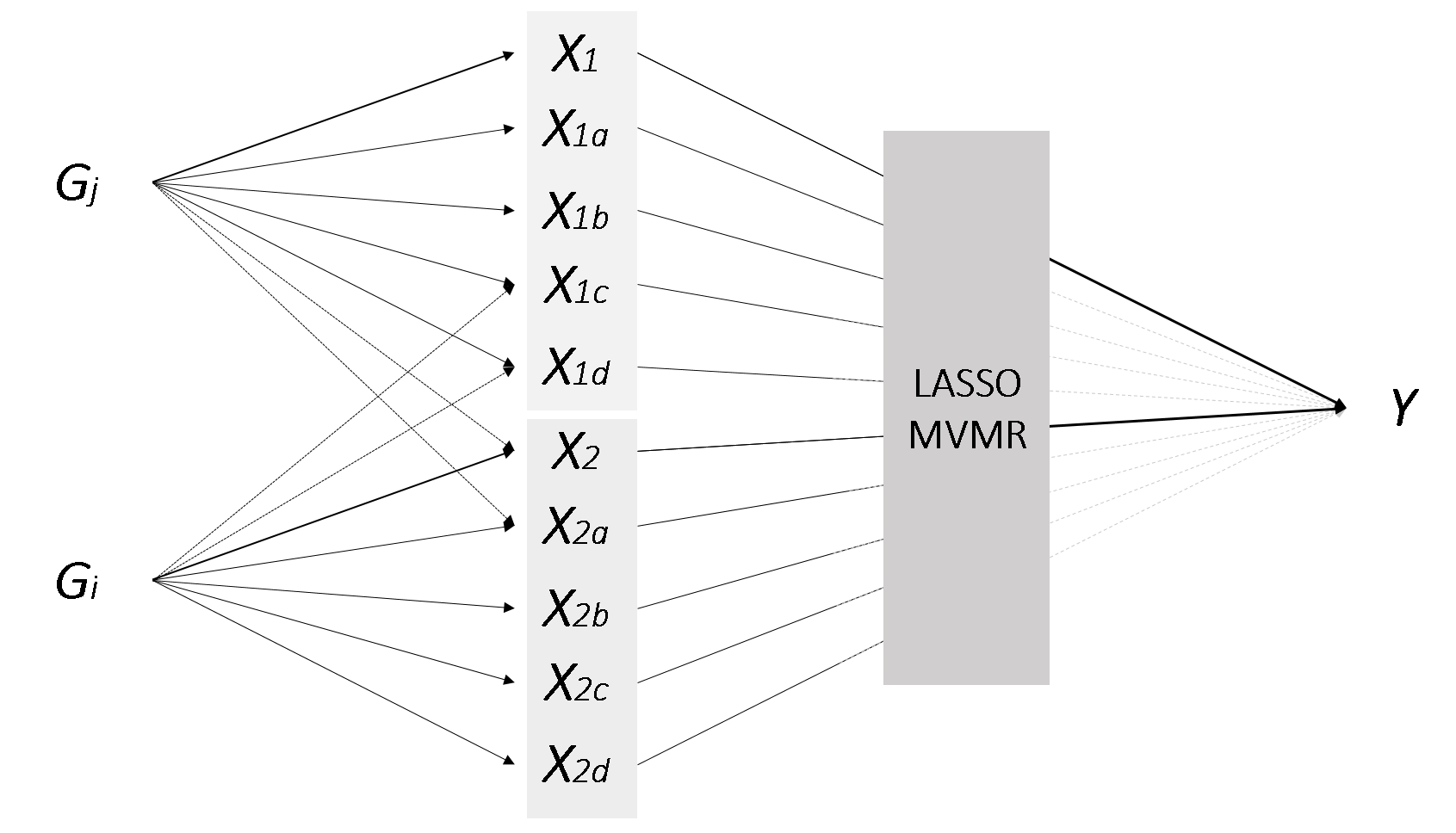
d. By selecting p based on lasso and using the lasso p->y estimates with the original standard errors

**Supplementary figure legends**

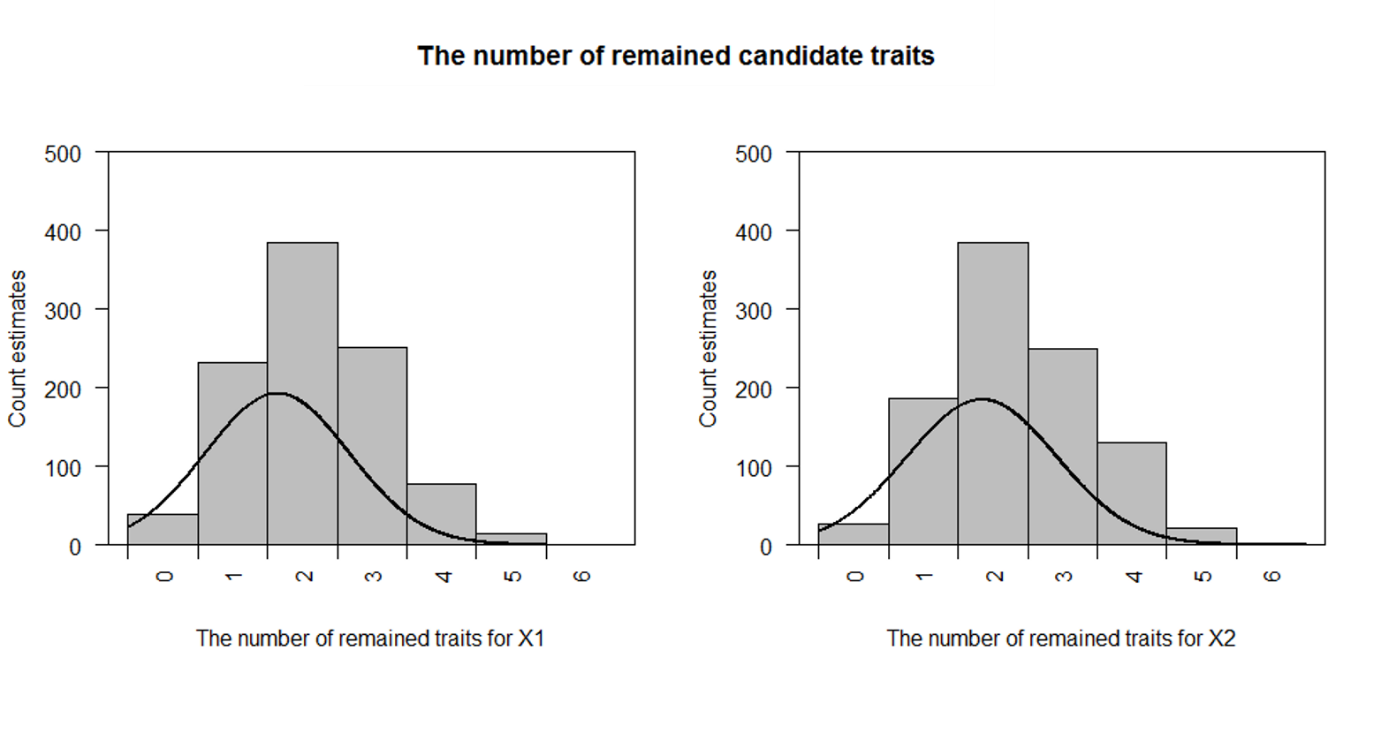
**Figure S1. The design of simulation 1.** Directed acyclic graphs illustrating that potentially account for the causal relationship between the candidate traits including the original trait and redundant traits (e.g. X1b, X2b, etc) and the outcome. The genetic variants Gj and Gi affect the traits. The dotted line between genetic variants and the traits indicates that genetic variant G affects the trait X1s also influences the traits X2s via horizontal pleiotropy. The LASSO step selects the traits that are associated with the outcome and have non-zero effects. The bold line between the traits and the outcome Y indicates the association between selected traits by the LASSO steps and the outcome.

**Figure S2. The number of candidate traits included in LASSO-MVMR.** The graph presents the number of traits that were survived after shrinkage step through the LASSO and were included in LASSO-MVMR analysis. The line indicates the normal distribution.

**Figure S1. The design of simulation 1.**

****

**Supplementary figure 2. The number of candidate traits included in LASSO-MVMR.**



**Supplementary Table S1. Sources of data used in this study.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Consortium/Studies** | **Use** | **Study population** | **Imputation** | **Model** | **Adjustments** |
| CARDIoGRAMplusC4D1 | SNP-Log odds CHD  SNP-Log odds MI | 22 233 cases and 64 762 controls of European descent  43 676 cases and 128,199 controls of European, South/East Asian, and Hispanic and African American ancestry | MACH, IMPUTE, or BIMBAM  (ref: HapMap II CEU)  IMPUTE, or MINIMAC  (ref: HapMap II/III or 1000 Genomes phase 1 ver. 3) | ADD | Age, sex and genomic control inflation factor |
| HaemGen | SNP-Platelet count2  SNP-Haemoglobin concentration3  SNP-Packed cell volume3  SNP-Red blood cell count3 | 66 867 individuals of European ancestry  71 861 individuals of European or South Asian ancestry  63 511 individuals of European or South Asian ancestry  66 214 individuals of European or South Asian ancestry | IMPUTE, BIMBAM, or MACH  (ref: HapMap II CEU)  MACH, IMPUTE, or BEAGLE  (ref: HapMap II) | ADD | Age, sex and other cohort‐specific covariates (where appropriate)  Cohort‐specific covariates  (where appropriate) |
| GUGC4 | SNP-Urate | 110 347 individuals of European ancestry | MACH, IMPUE, BEAGLE, BIMBAM  (ref: HapMap II CEU) | ADD | Age, sex, and study-specific covariates, if applicable |
| CKDGen5 | SNP-Serum Cystatin C | 133,413 individuals of European ancestry | MACH, IMPUTE, or BIMBAM  (ref: HapMap II CEU) | ADD | Age, sex, study site and genetic principal components (where appropriate). |
| IPSCSG6 | SNP-Primary sclerosing cholangitis | 4 796 cases and 19 955 population controls of European and American ancestry | IMPUTE2  (1000 Genomes Phase I ver. 3) | ADD |  |
| Dubois et al.7 | SNP-Celiac disease | 4 533 cases and 10 750 controls of European ancestry | BEAGLE and CEU, TSI, MEX and GIH  (ref: HapMap III) | ADD | Population stratification |
| GLGC8 | SNP-HDLC  SNP-LDLC  SNP-TC | 188 578 individuals of European, East Asian, South Asian, and African ancestry | MACH  (ref: HapMap II CEU) | ADD | Age, sex, principal components of  genomic ancestry (some studies),  and genomic control inflation factor |
| UK Biobank\* | SNP-SBP  SNP-Treatments  SNP-Anthropometric measures  SNP-Smoking  SNP-Alcohol drinking  SNP-Diagnosed diseases  SNP-Education  SNP-Exercise  SNP-Memory | Approx. 340 000 individuals of European ancestry | IMPUTE2  (ref: Haplotype Reference Consortium and UK10K haplotype resource) | ADD | Population structure (where appropriate). |

CARDIoGRAM, Coronary ARtery DIsease Genome-wide Replication and Meta-analysis; GUGC, Global Urate Genetics Consortium; CKDGen, Chronic Kidney Disease Genetics; IPSCSG, International Primary Sclerosing Cholangitis Study Group; GLGC, Global Lipids Genetics Consortium; CEU, Northern Europeans from Utah; ADD, additive.

\* UK Biobank GWAS results from Neale Lab: http://www.nealelab.is/blog/2017/7/19/rapid-gwas-of-thousands-of-phenotypes-for-337000-samples-in-the-uk-biobank.

**Supplementary Table S2. MR of candidate traits on each outcome (Excel).**

**Supplementary Table S3. Results of supplementary simulation 1.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Presence of horizontal pleiotropy** | | | |
|  | **No pleiotropy** | | **With pleiotropy** | |
|  | **Beta estimate** | **P value** | **Beta estimate** | **P value** |
| **LASSO-MVMR** |  |  |  |  |
| X1 | 0.195 (0.055) | < 1.00e-159 | 0.184 (0.056) | < 1.00e-159 |
| X2 | 0.192 (0.050) | < 1.00e-159 | 0.192 (0.044) | < 1.00e-159 |
| **MVMR** |  |  |  |  |
| X1 | 0.420 (0.493) | 1.09e-159 | 0.386 (0.605) | 1.73e-80 |
| X2 | -0.019 (0.494) | 8.89e-1 | 0.015 (0.605) | 4.29e- 1 |

S.D., standard deviation; MVMR, multivariable Mendelian randomization. 1 Beta estimates were derived from the LASSO-MVMR and normal MVMR. The simulate causal effect is 0.2.

**Supplementary Table S4. Results of supplementary simulation 2.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Probability to be selected by LASSO1** | | | |
| **Scenario2** | **S1** | **S2** | **S3** | **S4** |
| X1 | 1.000 | 0.467 | 1.000 | 0.225 |
| X2 | 0.473 | 1.000 | 1.000 | 0.217 |

S.D., standard deviation; MVMR, multivariable Mendelian randomization. 1 The values represent the probability of the number of times each trait is selected by LASSO out of 1000 repeated simulations. 2 X1 was instrumented by 100 genetic variants whilst X2 was instrumented by 20 genetic variants. 1 S1: X1 has an effect on Y (β=0.3), where X2 has no effect on Y (β=0.0); S2: X1 has no effect on Y (β=0.0), where X2 has an effect on Y (β=0.3); S3: Both of X1 and X2 have effects on Y (β=0.3); S4: Neither of X1 and X2 have effects on Y (β=0.0).

**References**

1. Nikpay M, Goel A, Won HH, et al. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet* 2015;47(10):1121-30. doi: 10.1038/ng.3396

2. Gieger C, Radhakrishnan A, Cvejic A, et al. New gene functions in megakaryopoiesis and platelet formation. *Nature* 2011;480(7376):201-8. doi: 10.1038/nature10659

3. van der Harst P, Zhang W, Mateo Leach I, et al. Seventy-five genetic loci influencing the human red blood cell. *Nature* 2012;492(7429):369-75. doi: 10.1038/nature11677

4. Huffman JE, Albrecht E, Teumer A, et al. Modulation of genetic associations with serum urate levels by body-mass-index in humans. *Plos One* 2015;10(3):e0119752. doi: 10.1371/journal.pone.0119752

5. Kottgen A, Pattaro C, Boger CA, et al. New loci associated with kidney function and chronic kidney disease. *Nat Genet* 2010;42(5):376-84. doi: 10.1038/ng.568

6. Ji SG, Juran BD, Mucha S, et al. Genome-wide association study of primary sclerosing cholangitis identifies new risk loci and quantifies the genetic relationship with inflammatory bowel disease. *Nat Genet* 2017;49(2):269-73. doi: 10.1038/ng.3745

7. Dubois PC, Trynka G, Franke L, et al. Multiple common variants for celiac disease influencing immune gene expression. *Nat Genet* 2010;42(4):295-302. doi: 10.1038/ng.543

8. Willer CJ, Schmidt EM, Sengupta S, et al. Discovery and refinement of loci associated with lipid levels. *Nat Genet* 2013;45(11):1274-83. doi: 10.1038/ng.2797