Bioinformatics, 31(6), 2015, 957–959
doi: 10.1093/bioinformatics/btu742
Advance Access Publication Date: 10 November 2014
Applications Note

OXFORD

Genome analysis

MeRP: a high-throughput pipeline for Mendelian randomization analysis

Peter Yin[1] and Benjamin F. Voight[2,3,*]

[1]Department of Biology, College of Arts and Sciences, University of Pennsylvania, Philadelphia, PA 19143, USA, [2]Department of Pharmacology and [3]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19143, USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on August 15, 2014; revised on November 3, 2014; accepted on November 4, 2014

Abstract

Summary: We present a Mendelian randomization (MR) pipeline (MeRP) to facilitate rapid, causal inference analysis through automating key steps in developing and analyzing genetic instruments obtained from publicly available data. Our tool uses the National Human Genome Research Institute catalog of associations to generate instrumental variable trait files and provides methods for filtering of potential confounding associations as well as linkage disequilibrium. MeRP generates estimated causal effect scores via a MR-score analysis using summary data for disease endpoints typically found in the public domain. We utilize our pipeline to develop genetic instruments for seven traits and evaluate potential causal relationships with two disease endpoints, observing two putatively causal associations between blood pressure and bone-mineral density with type 2 diabetes. Our tool emphasizes the importance of careful but systematic screening of large datasets for discovery and systematic follow-up.

Availability and implementation: MeRP is a free, open-source project and can be downloaded at http://github.com/py-merp/py-merp. Complete documentation can be found at http://py-merp.github.io. Requires Python 2.7, along with NumPy, SciPy.

Contact: bvoight@upenn.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The surge in genome-wide association studies (GWAS) has generated an abundance of single-nucleotide polymorphisms (SNPs) associated with a wide range of phenotypes, resulting in a rapidly growing body of research employing a statistical method utilizing SNPs as instrumental variables to draw causal inference between traits and diseases (Ebrahim and Davey-Smith, 2008; Hernán and Robins, 2006). This method, termed Mendelian randomization (MR), represents a genetic version of the randomized control trial where genetic variants are selected to correspond exclusively to a biomarker for which a causal estimate to a disease outcome is desired. The approach can address confounding by the fact alleles assort randomly at meiosis and also addresses reverse causality by the fact that disease states do not influence germline genetic variation. In 2013 alone, 115 MR-related studies were published. Thus, the limiting factor is no longer obtaining results from association studies but rather how we take advantage of these data efficiently and appropriately in MR studies today and in the future. As the number of GWAS and high-throughput sequencing studies extends the catalog of disease and trait associations, tools primed to take advantage of these data as they are produced will be essential.

The research community performing MR studies would be well advantaged by a computational pipeline that facilitates the use of the largest datasets available, supporting the labor-intensive process of curating data, streamlining the selection of genetic variants at an appropriate level of linkage disequilibrium (LD) with one another (if combined into a genetic score) and finally evaluated for association with confounding factors. Such a tool would generate hypotheses that

© The Author 2014. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

could help facilitate rapid transmission of discoveries to the research and ultimately, clinical communities, after sufficient evidence of causality has been accrued. To address these aims, we present MeRP, a computational pipeline to facilitate MR studies at high throughput.

2 Methods

The MeRP workflow consists of three components. First, the user obtains a catalog of SNP-Trait associations from public data to generate potential trait instrumental variable files (IVF). Then, a filtering algorithm is applied to these SNP-Trait files of interest to minimize associations with potential confounding factors and correlation across SNPs (due to LD), in order to satisfy MR conditions. Lastly, the user estimates a causal effect using the IVF for the given trait, along with a disease or endpoint dataset.

2.1 Obtaining genetic data from public domain

The National Human Genome Research Institute (NHGRI) maintains a compilation of data from GWAS publications consisting of 1961 publications and 14 012 SNPs (as of August 2014) (Welter *et al.*, 2014), and was used as our starting point. All SNP associations that are genome-wide significant ($P < 5 \times 10^{-8}$) are pooled by trait name into individual IVFs with information on the SNP id, trait effect allele, *P*-value and PubMed ID. We amend the alternative, non-effect allele for the genetic variant from the 1000 Genomes project (Phase I; 1000 Genomes Project Consortium *et al.*, 2012). Finally, we check for consistency in the direction of the units of change reported for trait in the NHGRI catalog, to ensure a uniform report and syntax across multiple SNPs for a single score (i.e. if the catalog stated the allele 'increases' the trait, we set the effect value on the trait to positive. Occasionally, the NHGRI catalog is not always consistent on this point).

2.2 Construction of potential genetic instruments

To help satisfy conditions of MR analysis, MeRP provides IVF filtering steps for LD and confounding trait associations. First, associations with potential confounding traits are assayed from within the NHGRI GWAS catalog, excluding associations with user-specified disease endpoints or traits that are intertwined with the trait of focus [e.g. low-density lipoprotein (LDL) with total cholesterol]. Next, users may specify a summary file containing association *P* values for interesting SNPs and potential confounders of interest for additional filtering. In our proof of concept, we provide a file with >2 million SNPs and 15 cardio/metabolic traits obtained from the public domain (see Supplementary Note). Users can filter SNPs that exceed a desired strength of association in these data and/or a specified threshold of weak associations. Users can also select a subset of *primary* confounding traits such that SNPs with a single association with one of these confounding traits are filtered out. Furthermore, users can ensure that the total number of potential confounding associations overall do not exceed a specified proportion of the IVF. Finally, remaining SNPs are optionally pruned for correlation by grouping them together based on a user-specified pairwise LD threshold via web query (see Supplementary Note), using the SNP with the most significant trait association value as the lead SNP for each LD group. The lead SNPs for each LD group comprise the list of genetic variants advanced into MR analysis.

2.3 Performing causal inference

Our pipeline implements one valid statistical method for estimating causal effects based on multi-SNP genetic instrument using summary data alone (Dastani, 2012). From this calculation, we obtain an estimate of the causal effect equivalent to that obtained from direct genotype MR under the assumptions that (i) the effect of individual SNPs are relatively weak, (ii) the SNPs used in IV are not correlated (no LD) and (iii) that the effects of multiple SNPs on the trait can be described as an additive, linear combination of the individual SNPs. These, in addition to satisfying the standard MR assumptions, provide an approach to estimate the causal effect and association statistics (Dastani, 2012). MeRP provides summary files for endpoint traits for the genetic instruments selected, allowing users to perform MR analysis with alternative statistics.

3 Results

3.1 Validation of computational pipeline

We applied MeRP to generate IVFs corresponding to seven different traits and estimated the causal effect score with coronary heart disease (CHD) in up to 22 233 cases and 64 762 controls, and type 2 diabetes (T2D) in up to 34 840 cases and 114 981 controls in two stages, both from publicly available data (see Supplementary Note). We began with trait–disease relationships for which there is evidence for and against causal effects: elevated LDL cholesterol (LDL-C) and systolic blood pressure (SBP) are causal risk factors for CHD, whereas genetic studies reject a causal, atheroprotective effect of elevated high-density lipoprotein (HDL) cholesterol (Voight *et al.*, 2012). For T2D, as elevated fasting glucose levels is a diagnostic criteria for the disease, it serves as a positive control. Figure 1 confirms these established relationships, with estimated effects consistent with previous studies. We also evaluated potential bias in SNP

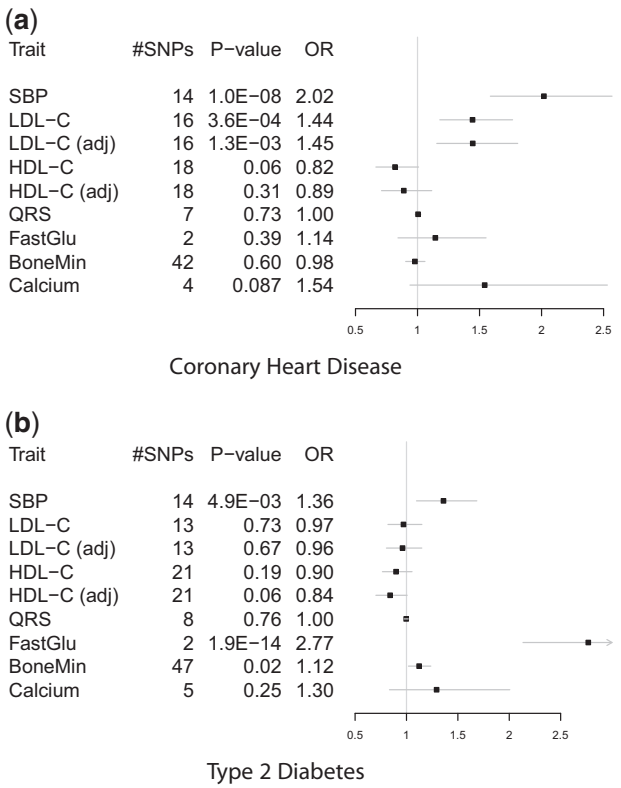

Fig.1. Estimated causal effects of traits with T2Ds and CHD. Odds ratio and 95% confidence intervals are shown in units of per SD increase except for fasting glucose (mmol/l), SBP (per 16 mmHg), QRS interval (ms) and calcium (mg/dl). Winner's curse adjusted results labeled as 'adj'

weights owing to Winner's curse (see Supplementary Note) and note that the HDL-CHD result seemed more consistent with the null after adjustment (Fig. 1).

3.2 Putatively novel findings

We observed causal association for elevated SBP ($P < 0.005$, passing a Bonferroni correction) and bone mineral density (BMD, $P < 0.02$) with T2D susceptibility. The positive SBP result could potentially result from biased ascertainment of CHD cases in the T2D study, as CHD is a known complication. However, to generate our observed odds ratio would require that >40% of our T2D cases to also be CHD cases, far higher than expected from random ascertainment. Another possibility could be that having high blood pressure leads to the use of blood pressure lowering medication, which itself increases the risk of T2D. It is known that hypertension predicts future diabetes, and ~70% of diabetic patients are hypertensive. Elliott and Meyer (2007) investigated antihypertensive therapies and their risks for T2D, noting a non-significant compared with placebo trending toward lower odds of T2D via blood pressure lowering through angiotensin receptor blockers, angiotensin-converting enzyme inhibitors or calcium channel blockers, but elevated risk for patients on beta-blockers or diuretics. Follow-up analyses will be required to assay the potential of these alternative mechanisms. Turning to BMD, studies have shown that T2D patients have elevated BMD compared with non-diabetic counterparts (Vestergaard, 2007). However, the mechanism of this association is unclear; one possibility is that BMD may serve as a correlate for another unmeasured biomarker. Our initial findings motivate work to refinement of the IVFs, investigation in prospective cohort data and clinical trials in order to further test these hypotheses.

4 Discussion

We present MeRP as a tool to help facilitate the labor-intensive steps in instrument creation for MR studies, and one valid, score-based analytical approach to summarize causal association. One caveat of the score-based approach is the concern of heterogeneity, as multiple genetic mechanisms or pathways may impact the trait of interest, and subsequently the outcome, differently. However, this hypothesis of heterogeneity is itself a testable one for multiple variants aggregated in a score, one that can be evaluated given the output from MeRP. With this output, MeRP also empowers users to investigate individual variants selected in their instrument, potentially applying prior biological knowledge to their inference, which may be critical to the longer-term success of such studies.

It is critical to emphasize that the evidence generated in an initial screen from MeRP adds but one piece of evidence to establish causality. An overwhelmingly compelling case for causality must involve further statistical analysis, studies in model systems, key insights into biological mechanisms and, ideally, evidence from interventional studies where applicable: a case built for LDL-C and CHD was done in precisely this way over the course of more than 100 years of directed research effort. Thus, we suggest that data-mining and filtering strategies provided by MeRP are themselves not to be taken lightly and serve as the first piece of evidence to focus on potentially meaningful relationships that undoubtedly require additional lines of evidence to provide casual evidence beyond a reasonable doubt.

Acknowledgements

P.Y. and B.F.V. thank the Alfred P. Sloan Foundation [BR2012-087], the American Heart Association [13SDG14330006] and the W.W. Smith Charitable Trust [H1201] who provided support for the work.

Conflict of interest: none declared.

References

1000 Genomes Project Consortium et al. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

Dastani,Z. *et al* (2012) Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals. *PLoS Genet*, **8**, e1002607.

Ebrahim,S. and Davey-Smith,G. (2008) Mendelian randomization: can genetic epidemiology help redress the failures of observational epidemiology? *Hum. Genet.*, **123**, 15–33.

Elliott,W.J. and Meyer,P.M. (2007) Incident diabetes in clinical trials of antihypertensive drugs: a network meta-analysis. *Lancet,* **369**, 201–207.

Hernán,M.A. and Robins,J.M. (2006) Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, **17**, 360–372.

Vestergaard,P. (2007) Discrepancies in bone mineral density and fracture risk in patients with type 1 and type 2 diabetes–a meta-analysis. *Osteoporos Int.*, **18**, 427–444.

Voight,B.F. *et al.* (2012) Plasma HDL cholesterol and risk of myocardial infarction: a Mendelian randomisation study. *Lancet*, **380**, 572–580.

Welter,D. *et al.* (2014) The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42**, 1001–1006.