Cold Spring Harbor Perspectives in Medicine
www.perspectivesinmedicine.org

Computational Tools for Causal Inference in Genetics

Tom G. Richardson, Jie Zheng, and Tom R. Gaunt

MRC Integrative Epidemiology Unit (IEU), Population Health Sciences, Bristol Medical School, University of Bristol, Bristol BS8 2BN, United Kingdom

Correspondence: Tom.G.Richardson@bristol.ac.uk

The advent of large-scale, phenotypically rich, and readily accessible data provides an unprecedented opportunity for epidemiologists, statistical geneticists, bioinformaticians, and also behavioral and social scientists to investigate the causes and consequences of disease. Computational tools and resources are an integral component of such endeavors, which will become increasingly important as these data continue to grow exponentially. In this review, we have provided an overview of computational software and databases that have been developed to assist with analyses in causal inference. This includes online tools that can be used to help generate hypotheses, publicly accessible resources that store summary-level information for millions of genetic markers, and computational approaches that can be used to leverage this wealth of data to study causal relationships.

Breakthroughs in genotyping arrays have propelled the discovery of genetic variants from across the human genome, which are robustly associated with complex traits and disease (Manolio 2010; Visscher et al. 2017; Trenkmann 2018). This widespread application of genome-wide association studies (GWAS) has, according to the GWAS catalog, identified over 150,000 associations between single base position genetic variants (known as single-nucleotide polymorphisms [SNPs]) based on the conventional threshold of $P < 5 \times 10^{-8}$ as of October 2019 (Buniello et al. 2019). Characterizing these findings can help develop insight into the hereditary component of complex disease, although, as discussed in the literature, they also provide an opportunity to undertake analyses in the field of causal inference (Davey Smith and Ebrahim 2003). However, identifying the best way to leverage and integrate the wealth of data from genetic association studies into analytical frameworks may be daunting for those without a background in bioinformatics.

In this article, we provide an overview of computational tools and resources that can be used to help disentangle causal relationships between risk factors and disease outcomes. We first consider web resources that have collated and stored the vast amounts of publicly accessible summary-level data for trait-associated genetic variants. Next, we provide an overview of resources that can be useful to help generate

Editors: George Davey Smith, Rebecca Richmond, and Jean-Baptiste Pingault
Additional Perspectives on Combining Human Genetics and Causal Inference to Understand Human Disease and Development available at www.perspectivesinmedicine.org

 doi: 10.1101/cshperspect.a039248
Cite this article as *Cold Spring Harb Perspect Med* 2021;11:a039248

hypotheses for causal inference studies. This includes techniques such as linkage disequilibrium (LD) score regression and constructing polygenic risk scores (PRSs). Furthermore, we highlight resources useful for variant annotation and prediction that can also have implications for causal analyses. This is followed by a discussion of the types of software that have been developed to harness these large-scale data and undertake studies in causal inference (such as Mendelian randomization [MR]). We also review some more recent developments in this paradigm, including the application of causal techniques to investigate molecular intermediate traits. We finish by showcasing resources that have precomputed millions of genetic and biological relationships. These may be valuable both in terms of hypothesis generation and identifying supporting evidence for existing research.

CATALOGS OF HARMONIZED SUMMARY STATISTICS FROM GENOME-WIDE ASSOCIATION STUDIES

The widespread adoption of GWAS can be attributed to technological advancements that have facilitated accurate and cost-effective approaches to genotyping large-scale populations (Oliphant et al. 2002; Syvänen 2005). However, the simplicity of GWAS both in terms of study design and interpretation of findings has also contributed to its popularity. Essentially, GWAS consist of millions of regressions where associations surviving multiple testing corrections cannot be attributed to reverse causation and are unlikely to be prone to confounding from unmeasured environmental factors. That said, differences to population structure (also known as "population stratification") may confound findings unless appropriate covariates are adjusted for. The overall premise of studies undertaking GWAS therefore remains the same, although there are differences in terms of how outcomes are derived/normalized, types of statistical models applied, genotyping platforms analyzed, and how to most appropriately account for population stratification.

As the number of GWAS being undertaken has increased exponentially in recent years, this has made it challenging to collate and curate a standardized database of SNP-trait summary statistics. This has motivated several research groups to develop web resources that store millions of harmonized results from GWAS. Perhaps the most well known is the GWAS catalog (www.ebi.ac.uk/gwas) hosted by the European Bioinformatics Institute (Buniello et al. 2019). Along with a web interface to query whether this database contains summary statistics for specific risk factors and outcomes, they have also made the full catalog accessible for download. Whereas most studies in the GWAS catalog originally only recorded genome-wide "significant" SNPs (commonly defined as SNPs with an association of $P < 5 \times 10^{-8}$), a growing number of studies are being made available with full summary statistics, with a dedicated application programming interface (API) to access the data programmatically (www.ebi.ac.uk/gwas/summary-statistics/docs). The European Bioinformatics Institute (EBI) GWAS catalog previously only recorded published GWAS, although it has recently begun accepting submission of unpublished datasets. Several unpublished GWAS datasets have been of tremendous help to the scientific community, such as those conducted using data from the UK Biobank (UKB) study (Sudlow et al. 2015; Bycroft et al. 2018) by the Neale laboratory (www.nealelab.is/uk-biobank).

A recently developed database of GWAS summary statistics is the GWAS Atlas (atlas.ctglab.nl) (Watanabe et al. 2019). This platform allows graphical illustrations of GWAS findings using Manhattan plots as well as other conventional downstream analyses undertaken by these types of studies (e.g., SNP heritability statistics [Yang et al. 2010], gene set enrichment analyses [de Leeuw et al. 2015]). SNPs and gene-based phenome-wide association studies (PheWAS) evaluations can also be undertaken, which involves collating associations at a single genetic locus across many different outcomes and traits. Other noteworthy platforms include PhenoScanner (www.phenoscanner.medschl.cam.ac.uk) (Staley et al. 2016) and the GeneATLAS (geneatlas.roslin.ed.ac.uk) (Canela-Xandri et al. 2018).

Many large-scale GWAS consortia and studies have contributed substantially to these databases, which would not be possible without their dedication to open science. Consequently, there is also a large degree of overlap between the summary statistics hosted by each of these platforms.

A database that has been developed more specifically with the aim of facilitating endeavors in causal inference is the Integrative Epidemiology Unit (IEU) OpenGWAS database (gwas.mrcieu.ac.uk). This database is the primary data source for the MR-Base platform (www.mrbase.org) (Hemani et al. 2018), which will be discussed later in this review. Similar to the GWAS catalog, the IEU OpenGWAS database has an API (gwasapi.mrcieu.ac.uk), which retrieves relevant GWAS summary statistics based on queries from users. Effect estimates for an SNP across all traits in the database can also be undertaken at gwas.mrcieu.ac.uk/phewas. An overview of the resources highlighted in this section along with a glossary of terms can be found in Table 1. A flowchart of how summary statistics obtained from these databases can be used for causal inference analyses is shown in Figure 1, to accompany the subsequent sections of this article.

GENERATING HYPOTHESES FOR CAUSAL INFERENCE USING INFORMATIC APPROACHES

Various approaches have been used in the fields of genetics and bioinformatics to uncover potential causal relationships between environmental exposures and disease outcomes. Such analyses can therefore be valuable in terms of generating hypotheses, for which more rigorous evaluations can then be undertaken to more robustly discern whether they may be attributed to an underlying causal effect (Pingault et al. 2018). An increasingly popular approach to this in the field of genetic epidemiology is genetic correlation analyses, typically undertaken using LD score regression (Bulik-Sullivan et al. 2015a,b). If two traits share a high genetic correlation, then it suggests that a substantial proportion of genes contribute to the variation in both. This may therefore indicate a possible causal relationship (i.e., an individual's genotype predisposes them to trait A, which in turn has a causal effect on trait B, also referred to as "vertical pleiotropy") (Fig. 2A). Conversely, such correlations could arise by a shared genetic etiology (i.e., the genes that predispose an individual to trait A also happen to influence varia-

Table 1. An overview of resources containing genetic associations that can be harnessed to infer causal relationships along with a glossary of terms used in this field

Resource	URLs
European Bioinformatics Institute genome-wide association studies (EBI GWAS) catalog	www.ebi.ac.uk/gwas
The UK Biobank (UKB) study	www.ukbiobank.ac.uk
The Neale laboratory UKB GWAS analysis	www.nealelab.is/uk-biobank
The GWAS Atlas	atlas.ctglab.nl
PhenoScanner	www.phenoscanner.medschl.cam.ac.uk
GeneATLAS	geneatlas.roslin.ed.ac.uk
Integrative Epidemiology Unit (IEU) GWAS database	gwas.mrcieu.ac.uk
MR-Base	www.mrbase.org
Glossary of terms	**Definitions**
Population stratification	Differences in the allele frequencies of genetic variants between populations
Heritability	Variation in a complex trait or disease outcome attributed to genetic variation in a population
Gene set enrichment	An overrepresentation of a predetermined group of genes compared to a randomly selected background set

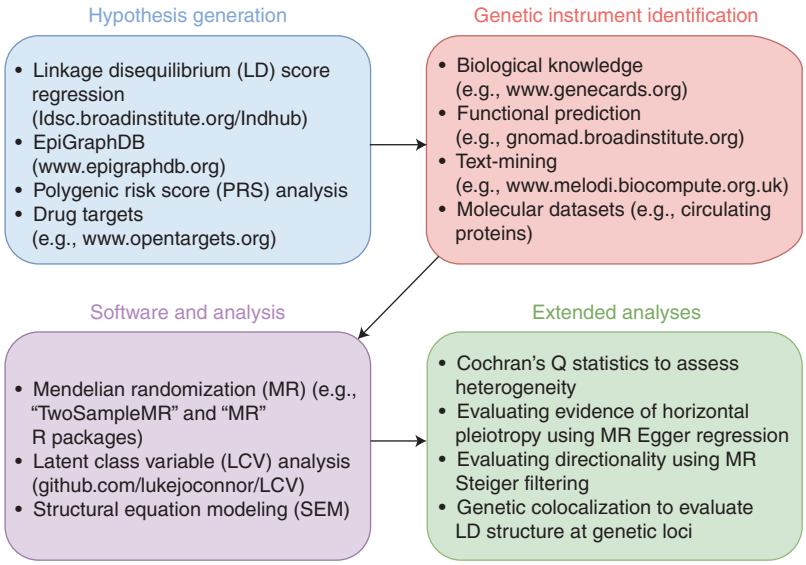

Figure 1. A flowchart showing potential analysis pipelines that can be constructed using the tools and resources described in this article.

tion in trait B but via a separate causal pathway, also known as "horizontal pleiotropy") (Fig. 2B).

Along with standalone software, which is available to undertake LD score regression (github.com/bulik/ldsc), a database full of precomputed correlations has been developed known as LD Hub (ldsc.broadinstitute.org/ldhub) (Zheng et al. 2017). Furthermore, LD Hub allows users to upload their own GWAS summary statistics to screen for genetic correlations using publicly available results. Although these types of analyses have become routinely undertaken using summary statistics, it should be noted that they can be more powerfully estimated using individual-level data (Lee et al. 2012; Speed et al. 2017).

Another area of increasing interest in the field of genetic epidemiology in recent years is the use of PRS to predict risk of disease. A PRS is commonly defined as the summed score of risk alleles that an individual harbors, weighted by effect sizes obtained from GWAS. Although this concept has existed for some time (Evans et al. 2013), PRSs are now being considered to be "coming of age" as they are being constructed in increasingly large sample sizes thanks to cohorts with accessible individual-level data (e.g., the UKB study) (Khera et al. 2018). Although conventionally used for predicting the same outcome they have been constructed for (e.g., assessing how well a PRS for coronary heart disease [CHD] predicts incidence of it in a separate population), their application to different outcomes may potentially highlight underlying causal relationships (Evans et al. 2013). For example, the PRS of low-density lipoproteins (LDLs) predicts CHD very strongly in a separate population because LDL is a well-known risk factor for this outcome (Richardson et al. 2019a). However, as with LD score regression, associations between a PRS and an alternate trait could also indicate shared genetic etiology instead of causality. Such findings should therefore also be rigorously evaluated using follow-up analyses to try and distinguish between these two explanations.

A popular approach to construct PRS is by using the genetic association software PLINK (Chang et al. 2015), although other bespoke PRS software options are also available such as PRSice (Euesden et al. 2015). Finally, a precomputed database of PRS associations generated within the UKB study can be found at mrcieu.mrsoftware.org/PRS_atlas, which allows interactive visualizations of findings in a phenome-wide manner (e.g., assessing the systematic

Cold Spring Harbor Perspectives in Medicine
www.perspectivesinmedicine.org

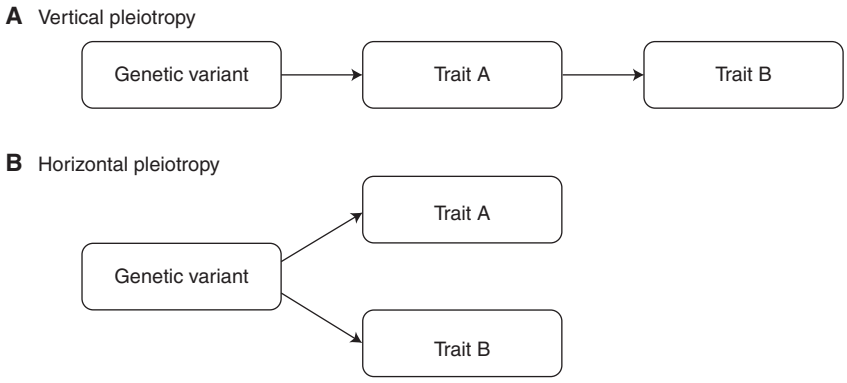

Figure 2. Direct acyclic graphs depicting the difference between (*A*) vertical pleiotropy, and (*B*) horizontal pleiotropy.

association between a selected PRS and 551 different traits) (Fig. 3; Richardson et al. 2019a).

CURATED RESOURCES TO HELP DESIGN CAUSAL ANALYSES

Along with the data-driven approaches highlighted in the previous section, which can help develop hypotheses, more conventional approaches, such as reviewing the relevant literature, remain integral to study designs in causal inference using genetic data. However, given that this is typically a time-consuming endeavor, there are curated databases that can assist in this regard. For instance, developing insight into the genetic variants being harnessed (or the region of the genome they are located at) can be gathered using resources such as the Online Mendelian Inheritance in Man (OMIM) (omim.org) (Amberger et al. 2015), GeneCards (www.genecards.org) (Stelzer et al. 2016), and Ensembl (ensembl.org) (Zerbino et al. 2018) platforms. Text-mining is another branch of bioinformatics that can be valuable for hypothesis generation. One such approach to this is MELODI (www.melodi.biocompute.org.uk), which can help identify links and potential intermediates between risk factors and outcomes, as well as highlight genes that may play a role in conferring disease risk (Elsworth et al. 2018). For example, a user may have a target disease outcome they are interested in researching the causal determinants of. Platforms such as MELODI allow users to mine the relevant literature for that disease and return a list of potential risk factors for it, which can be followed up in causal analysis.

Biological insight into the genes that underlie an association from GWAS can help identify genetic variants useful as powerful instrumental variables for causal analyses. This is because they can help to validate assumptions being made, such as those regarding horizontal pleiotropy as previously mentioned, or that genetic variants only influence a disease outcome because of their initial effect on the risk factor being studied (Davey Smith and Ebrahim 2003). For example, SNPs at the *CHRNA5* locus have been used previously to instrument smoking as a risk factor, given that this gene is a nicotinic acetylcholine receptor and plays an important role in smoking heaviness and cessation (Taylor et al. 2014; Gage et al. 2017). Likewise, the variant rs671 has been used to instrument alcohol intake in previous studies. This missense SNP (located in the *ALDH2* gene) is responsible for encoding a form of the aldehyde dehydrogenase 2 protein, which does not effectively metabolize alcohol. As a consequence, individuals who carry two copies of the minor allele at this locus typically do not drink as they are less able to clear alcohol from their system. As such, it has been used in Asian populations (where it is polymorphic) to demonstrate the effects of alcohol on health and disease, such as increased risk of stroke and high blood pressure (Chen et al. 2008; Cho et al. 2015; Millwood

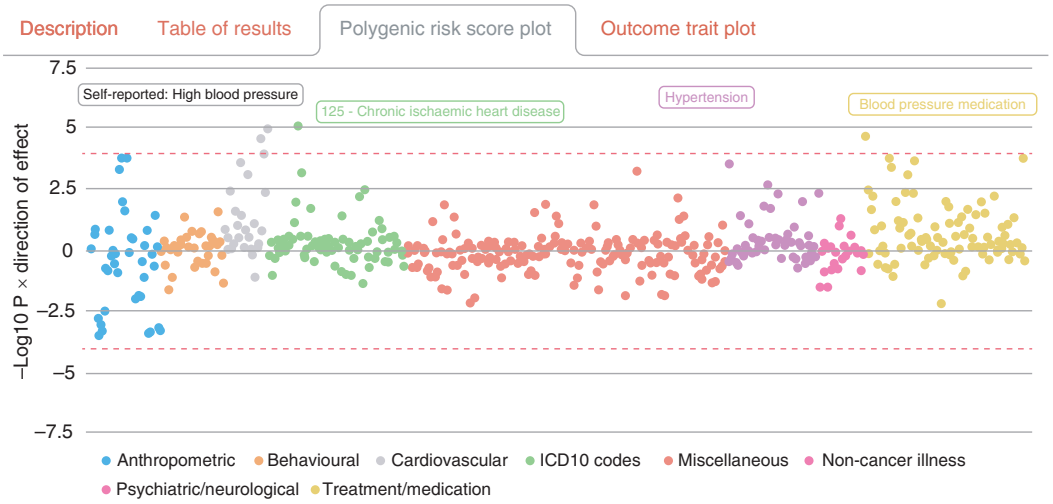

Figure 3. The atlas of polygenic risk score (PRS) associations at mrcieu.mrsoftware.org/PRS_atlas can be helpful in terms of generating hypotheses for future analyses in causal inference. For example, the plot displayed here represents the association between the ischemic stroke PRS and 551 outcomes. The only points that survive multiple testing here (indicated by the dashed red line) represent outcomes related to blood pressure and the ICD-10 for ischemic heart disease. This is likely due to blood pressure being an established risk factor for heart disease, as well as stroke and heart disease having a shared etiology. Similar lookups can be performed for all 162 PRSs in this atlas.

et al. 2019). This illustrates the value of selecting genetic variants as instrumental variables with prior biological knowledge for causal analyses, particularly as GWAS continue to uncover increasingly large number of variants to select instruments from.

Along with biological relevance, the predicted consequence of genetic variants may also be important to consider for future studies. For example, a genetic variant in a coding sequence, which is predicted to change the amino acid sequence of a protein (known as a "non-synonymous" variant), is more likely to have a downstream consequence on a risk factor or disease as opposed to SNPs predicted to have a benign impact. In terms of software, the gnomAD database (gnomad.broadinstitute.org) is regarded as having one of the most comprehensive catalogs of predicted variant consequences to date (Karczewski et al. 2019). There are also variants that are predicted to result in a protein losing its function (known as "loss-of-function" variants). The majority of these are rare in populations because of their detrimental impact, although particularly given the increasing sample sizes for exome sequencing datasets, they may be potentially useful for future causal analyses. Other resources that provide information regarding variant consequences include ClinVar (www.ncbi.nlm.nih.gov/clinvar), which is based on data collection from clinical testing, research, and reports from the literature (Landrum and Kattman 2018). Additionally, there are various machine learning approaches that base their predictions of variant consequences on biological features from a plethora of resources. This includes FATHMM (fathmm.biocompute.org.uk) (Shihab et al. 2013), CADD (cadd.gs.washington.edu) (Kircher et al. 2014), SIFT (sift.bii.a-star.edu.sg) (Ng and Henikoff 2003), and Polyphen-2 (genetics.bwh.harvard.edu/pph2) (Adzhubei et al. 2010).

SOFTWARE IN CAUSAL INFERENCE

The resources reviewed in the previous sections provide a treasure trove of data for researchers interested in causal inference, given the scale of SNP-trait effect estimates concerning a vast breadth of risk factors and outcomes. There is

also a plethora of software available to leverage GWAS results using these platforms to develop insight into causal relationships in disease. For instance, the IEU OpenGWAS database previously mentioned is specifically tailored toward MR analyses, a method by which genetic variants are harnessed as instrumental variables to help infer causality among correlated traits. The TwoSampleMR package (github.com/MRCIEU/TwoSampleMR) (Hemani et al. 2018) has been developed to contact the API for this database and provides a convenient way to extract and use the data for two-sample MR. A large proportion of the GWAS results contained in the IEU GWAS database have been gathered from analyses in the UKB study (Elsworth et al. 2020). This summary-level data has also been standardized to facilitate the interpretation of results from MR analyses. A user-friendly front-end for this software is also available at www.mrbase.org.

The TwoSampleMR package includes various techniques developed in the field to evaluate causal relationships using genetic variants. This includes (but is not limited to) established methods in the field such as the Wald ratio, inverse variance weighted (Burgess et al. 2013), MR-Egger (Bowden et al. 2015), weighted median (Bowden et al. 2016), and weighted mode (Hartwig et al. 2017) approaches. There are also various sensitivity analyses commonly used in MR analyses, which are facilitated by the package, such as leave-one-out analyses, evaluations of horizontal pleiotropy based on the MR-Egger regression intercept, and analysis of heterogeneity in a two-sample setting. Plots are also readily generated by the platform to assist with such evaluations. Increasingly important is identifying and removing instruments in analyses that are prone to invalidating the underlying assumptions of MR. This includes SNPs, which influence a disease outcome along a different causal pathway to one involving a given exposure. This is the motivation behind the MR Steiger directionality test (Hemani et al. 2017b), also referred to as "Steiger filtering." This sensitivity analysis involves removing SNPs from MR analyses, which are more strongly correlated with the disease outcome than the exposure being analyzed and can also be implemented using the TwoSampleMR package.

This software also has the functionality to convert genetic data into a suitable format for other packages. This includes the "MendelianRandomization" package (Yavorska and Burgess 2017). Other standalone R packages developed in the field are also available, such as MR-PRESSO (Verbanck et al. 2018), MR-RAPS (Zhao et al. 2019), and MR-TRYX (Cho et al. 2020). Additionally, power calculations for MR studies can be undertaken using an online tool accessible at cnsgenom ics.com/shiny/mRnd (Brion et al. 2013).

There are also recently developed software approaches that harness genetic variants to infer causal relationships and do not rely upon the assumptions of MR. This includes a latent class variable (LCV) approach based on the genetic correlation between two traits (O'Connor and Price 2018) with open source code available at github.com/lukejoconnor/LCV. Additionally, approaches using structural equation modeling (SEM) are becoming increasingly popular in the field of genetics to help infer causality (Warrington et al. 2018), and the software to undertake such analyses is being developed in parallel (for example, at github.com/MichelNivard/GenomicSEM/wiki) (Grotzinger et al. 2019).

CAUSAL VARIANTS AND INTEGRATION WITH MOLECULAR TRAITS

LD structure across the human genome results in nearby SNPs being correlated with one another. Consequently, it can be challenging to pinpoint the exact SNP that is responsible for an association with a complex trait or disease. Furthermore, this correlation that can exist between neighboring genetic variants also makes it challenging to disentangle independent instruments to use in a causal framework. Using correlated instruments in an MR analysis may lead to false-positive discoveries, essentially because of "double counting." A popular approach to separate independent effects is using genetic data from a separate population of individuals (known as a "reference panel") and undertake "LD clumping." This involves identifying the

SNP with the strongest association in a region (typically based on observed *P*-values) and then removing all nearby effects that are in LD with the lead SNP. LD clumping can be routinely undertaken using the software PLINK (Chang et al. 2015) and can be undertaken using summary-level information (as well as with individual-level data) as long as an appropriate reference panel based on a population of the same ancestry as analyzed is used. Additionally, using conditional analyses to identify independent SNPs, for instance using the GCTA-mtCOJO software (Zhu et al. 2018), is becoming increasingly popular as an alternative to LD clumping.

LD structure also makes it difficult to distinguish whether overlapping association signals from separate GWAS are driven by the same causal SNP. This can be directly relevant to causal analyses as two traits that share the same causal variant at a locus suggests that a relationship may exist between them (e.g., an SNP has a direct effect on trait A, which in turn influences trait B). Methods that examine whether multiple traits share a causal variant at a region are known as techniques in "genetic colocalization." Popular approaches include eCAVIAR (Hormozdiari et al. 2016), coloc (Giambartolomei et al. 2014), moloc (Giambartolomei et al. 2018), enloc (Wen et al. 2017), and Heidi (Zhu et al. 2016). In particular, their use is becoming popular to investigate whether GWAS associations overlap with effects on molecular traits, such as gene expression (Zhu et al. 2016; Taylor et al. 2019; Porcu et al. 2020), epigenetic signatures (e.g., DNA methylation) (Hatcher et al. 2019; Richardson et al. 2019b), and plasma proteins (McGowan et al. 2019; Zheng et al. 2020). Identifying evidence suggesting that a variant influences a molecular trait as well as a disease outcome has various translational applications. For instance, such findings may help understand which tissue types are important for a given disease outcome (e.g., genes that influence our risk of neurological disease likely exert their effect through changes to expression in brain tissue) (Franceschini et al. 2018; Taylor et al. 2019). Likewise, techniques in genetic colocalization are becoming increasingly used to prioritize which genes may make worthwhile drug targets (Kibinge et al. 2020). Techniques in genetic colocalization require information regarding the LD structure at regions analyzed, which can either be derived from individual-level data or a valid reference panel. Current developments in this paradigm involve sensitivity analyses concerning the prior distributions of two traits sharing a causal variant, as well as undertaking conditional analyses on lead variants in a region (Wallace 2020).

COMPUTATIONAL RESOURCES TO SYSTEMATICALLY EVALUATE EVIDENCE OF CAUSALITY

Another trend in the field is the increasing popularity of web resources that allow users to query databases on precomputed results from causal analyses. This includes the EpiGraphDB platform (epigraphdb.org), which has a dedicated API and a standalone R package (github.com/MRCIEU/epigraphdb-r). EpiGraphDB contains results from MR-EvE and also from systematic MR of protein and transcript levels on outcomes from the IEU GWAS database. The systematic results of MR analysis of genes and proteins on hundreds of diverse traits can be a powerful approach to evaluate potential beneficial and adverse effects of therapeutic intervention (Richardson et al. 2020; Zheng et al. 2020). The EpiGraphDB platform has multiple browsers, which allow these results to be interrogated, such as the genetically predicted effects of 1740 plasma protein (epigraphdb.org/pqtl) and 16,058 whole-blood-derived transcripts (epigraphdb.org/xqtl) on 576 complex traits and outcomes. In addition to causal estimates, EpiGraphDB integrates data from the biomedical literature, pathway databases, drug target databases, observational correlations, and genetic correlations. The integration of systematic molecular MR results with pathway and drug target data offers valuable insights for drug target prioritization, validation, and repositioning purposes (see Fig. 4).

EpiGraphDB also hosts the MR of "everything vs everything" (MR-EvE) database, which contains effect estimates from MR analyses to evaluate the pairwise relationships between 2407 traits and diseases (Hemani et al. 2017a). This was achieved through the use of a mixture-

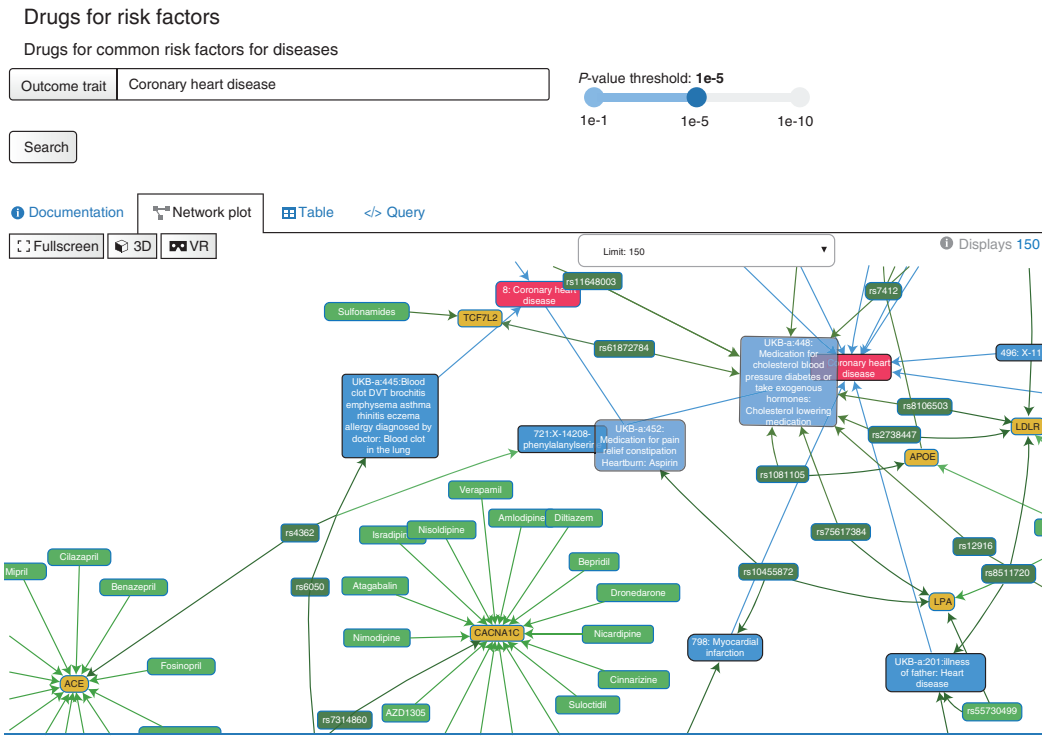

Figure 4. The EpiGraphDB platform (epigraphdb.org) hosts a treasure trove of information that can assist with causal inference analyses. For instance, displayed in this figure are various environmental factors and genes that provide evidence of a causal effect on risk of coronary heart disease (CHD). Furthermore, depicting how these factors may collectively contribute to disease risk as shown using automatically generated network graphs allows the user to identify supporting evidence and hypothesis generation.

of-experts machine learning framework to predict the most appropriate MR method to undertake analyses between a pair of phenotypes. Furthermore, information on this can be found in the associated preprint (Hemani et al. 2017a). Other resources that have been developed for this purpose include the Open Targets platform (www.opentargets.org), which has been constructed by a partnership of seven academic and industrial institutions with the overall aim of prioritizing the efficacy and safety of therapeutic targets (Koscielny et al. 2017).

CONCLUSIONS AND FUTURE DIRECTIONS

The availability of high-dimensional epidemiological datasets is increasing at an exponential rate both in terms of scale and complexity. There is therefore a demand for computational tools and resources that can collate and analyze these data appropriately to assist researchers in the field of causal inference. Key developments that are likely needed by future research include an emphasis on how to integrate various lines of data and evidence together. Furthermore, there are other factors that are becoming increasingly important to account for when using GWAS results to investigate causality. These include assortative mating (Howe et al. 2019) and dynastic effects (Brumpton et al. 2020), the latter of which is discussed recently in a review of MR in family study designs (Hwang et al. 2020).

Along with statistical and methodological considerations, novel computational methods and resources will be needed to rise to these forthcoming challenges. Doing so will improve our capability to make robust claims of causality through the triangulation of many lines of evi-

Cold Spring Harbor Perspectives in Medicine
www.perspectivesinmedicine.org

dence (Munafò and Davey Smith 2018; Munafò 2020). The characterization of increasingly complex causal pathways is another aspect of future research where computational approaches will play an important role. There will need to be additional consideration in terms of the computational burden, which will be incurred from analyzing large-scale and complex datasets. As highlighted in this review, many approaches are beginning to be developed to handle summary-level information as well as individual-level data. This will be increasingly important for future research, particularly given that individual-level datasets will continue to grow exponentially in terms of scale. Challenges such as this present an exciting opportunity for the development of cutting-edge research in the fields of bioinformatics and causal inference.

ACKNOWLEDGMENTS

This work was supported by the Integrative Epidemiology Unit, which receives funding from the UK Medical Research Council and the University of Bristol (MC_UU_00011/4). T.R.G conducts research at the NIHR Biomedical Research Centre at the University Hospitals Bristol NHS Foundation Trust and the University of Bristol. The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research, or the Department of Health. T.G.R is a UKRI Innovation Research Fellow (MR/S003886/1).

REFERENCES

*Reference is also in this collection

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nat Methods* **7:** 248–249. doi:10.1038/nmeth0410-248

Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. 2015. OMIM.org: Online Mendelian Inheritance in Man (OMIM), an online catalog of human genes and genetic disorders. *Nucleic Acids Res* **43:** D789–D798. doi:10.1093/nar/gku1205

Bowden J, Davey Smith G, Burgess S. 2015. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol* **44:** 512–525. doi:10.1093/ije/dyv080

Bowden J, Davey Smith G, Haycock PC, Burgess S. 2016. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genet Epidemiol* **40:** 304–314. doi:10.1002/gepi.21965

Brion MJ, Shakhbazov K, Visscher PM. 2013. Calculating statistical power in Mendelian randomization studies. *Int J Epidemiol* **42:** 1497–1501. doi:10.1093/ije/dyt179

Brumpton B, Sanderson E, Heilbron K, Hartwig FP, Harrison S, Vie GÅ, Cho Y, Howe LD, Hughes A, Boomsma DI, et al. 2020. Avoiding dynastic, assortative mating, and population stratification biases in Mendelian randomization through within-family analyses. *Nat Commun* **11:** 3519. doi:10.1038/s41467-020-17117-4

Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR; ReproGen Consortium; Psychiatric Genomics Consortium; Genetic Consortium for Anorexia Nervosa of The Wellcome Trust Case Control Consortium; Duncan L, et al. 2015a. An atlas of genetic correlations across human diseases and traits. *Nat Genet* **47:** 1236–1241. doi:10.1038/ng.3406

Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J; Schizophrenia Working Group of The Psychiatric Genomics Consortium; Patterson N, Daly MJ, Price AL, Neale BM. 2015b. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47:** 291–295. doi:10.1038/ng.3211

Buniello A, Macarthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, et al. 2019. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47:** D1005–D1012. doi:10.1093/nar/gky1120

Burgess S, Butterworth A, Thompson SG. 2013. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet Epidemiol* **37:** 658–665. doi:10.1002/gepi.21758

Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, et al. 2018. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562:** 203–209. doi:10.1038/s41586-018-0579-z

Canela-Xandri O, Rawlik K, Tenesa A. 2018. An atlas of genetic associations in UK Biobank. *Nat Genet* **50:** 1593–1599. doi:10.1038/s41588-018-0248-z

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4:** 7. doi:10.1186/s13742-015-0047-8

Chen L, Smith GD, Harbord RM, Lewis SJ. 2008. Alcohol intake and blood pressure: a systematic review implementing a Mendelian randomization approach. *PLoS Med* **5:** e52. doi:10.1371/journal.pmed.0050052

Cho Y, Shin SY, Won S, Relton CL, Davey Smith G, Shin MJ. 2015. Alcohol intake and cardiovascular risk factors: a Mendelian randomisation study. *Sci Rep* **5:** 18422. doi:10.1038/srep18422

Cho Y, Haycock PC, Sanderson E, Gaunt TR, Zheng J, Morris AP, Davey Smith G, Hemani G. 2020. Exploiting horizontal pleiotropy to search for causal pathways within a Mendelian randomization framework. *Nat Commun* **11:** 1010. doi:10.1038/s41467-020-14452-4

Davey Smith G, Ebrahim S. 2003. "Mendelian randomization": can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* **32:** 1–22. doi:10.1093/ije/dyg070

De Leeuw CA, Mooij JM, Heskes T, Posthuma D. 2015. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol* **11:** e1004219. doi:10.1371/journal.pcbi.1004219

Elsworth B, Dawe K, Vincent EE, Langdon R, Lynch BM, Martin RM, Relton C, Higgins JPT, Gaunt TR. 2018. MELODI: mining enriched literature objects to derive intermediates. *Int J Epidemiol* **47:** 369–379 doi:10.1093/ije/dyx251

Elsworth B, Lyon M, Alexander T, Liu Y, Matthews P, Hallett J, Bates P, Palmer T, Haberland V, Davey Smith G, et al. 2020. The MRC IEU OpenGWAS data infrastructure. bioRxiv doi:10.1101/2020.08.10.244293

Euesden J, Lewis CM, O'Reilly PF. 2015. PRSice: polygenic risk score software. *Bioinformatics* **31:** 1466–1468. doi:10.1093/bioinformatics/btu848

Evans DM, Brion MJ, Paternoster L, Kemp JP, McMahon G, Munafò M, Whitfield JB, Medland SE, Montgomery GW; The GIANT Consortium; et al. 2013. Mining the human phenome using allelic scores that index biological intermediates. *PLoS Genet* **9:** e1003919. doi:10.1371/journal.pgen.1003919

Franceschini N, Giambartolomei C, De Vries PS, Finan C, Bis JC, Huntley RP, Lovering RC, Tajuddin SM, Winkler TW, Graff M, et al. 2018. GWAS and colocalization analyses implicate carotid intima-media thickness and carotid plaque loci in cardiovascular outcomes. *Nat Commun* **9:** 5141. doi:10.1038/s41467-018-07340-5

Gage SH, Jones HJ, Taylor AE, Burgess S, Zammit S, Munafò MR. 2017. Investigating causality in associations between smoking initiation and schizophrenia using Mendelian randomization. *Sci Rep* **7:** 40653. doi:10.1038/srep40653

Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, Plagnol V. 2014. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* **10:** e1004383. doi:10.1371/journal.pgen.1004383

Giambartolomei C, Zhenli Liu J, Zhang W, Hauberg M, Shi H, Boocock J, Pickrell J, Jaffe AE; The CommonMind Consortium; Pasaniuc B, et al. 2018. A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics* **34:** 2538–2545. doi:10.1093/bioinformatics/bty147

Grotzinger AD, Rhemtulla M, De Vlaming R, Ritchie SJ, Mallard TT, Hill WD, Ip HF, Marioni RE, McIntosh AM, Deary IJ, et al. 2019. Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat Hum Behav* **3:** 513–525. doi:10.1038/s41562-019-0566-x

Hartwig FP, Davey Smith G, Bowden J. 2017. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int J Epidemiol* **46:** 1985–1998. doi:10.1093/ije/dyx102

Hatcher C, Relton CL, Gaunt TR, Richardson TG. 2019. Leveraging brain cortex-derived molecular data to elucidate epigenetic and transcriptomic drivers of complex traits and disease. *Transl Psychiatry* **9:** 105. doi:10.1038/s41398-019-0437-2

Hemani G, Bowden J, Haycock P, Zheng J, Davis O, Flach P, Gaunt T, Smith GD. 2017a. Automating Mendelian randomization through machine learning to construct a putative causal map of the human phenome. bioRxiv 173682.

Hemani G, Tilling K, Davey Smith G. 2017b. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS Genet* **13:** e1007081. doi:10.1371/journal.pgen.1007081

Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, Laurin C, Burgess S, Bowden J, Langdon R, et al. 2018. The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **7:** e34408. doi:10.7554/eLife.34408

Hormozdiari F, Van De Bunt M, Segrè AV, Li X, Joo JWJ, Bilow M, Sul JH, Sankararaman S, Pasaniuc B, Eskin E. 2016. Colocalization of GWAS and eQTL signals detects target genes. *Am J Hum Genet* **99:** 1245–1260. doi:10.1016/j.ajhg.2016.10.003

Howe LJ, Lawson DJ, Davies NM, St. Pourcain B, Lewis SJ, Davey Smith G, Hemani G. 2019. Genetic evidence for assortative mating on alcohol consumption in the UK Biobank. *Nat Commun* **10:** 5039. doi:10.1038/s41467-019-12424-x

* Hwang L-D, Davies NM, Warrington NM, Evans DM. 2020. Integrating family-based and Mendelian randomization designs. *Cold Spring Harb Perspect Med* doi:10.1101/cshperspect.a039503

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, et al. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581:** 434–443. doi:10.1038/s41586-020-2308-7

Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, Natarajan P, Lander ES, Lubitz SA, Ellinor PT, et al. 2018. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* **50:** 1219–1224. doi:10.1038/s41588-018-0183-z

Kibinge NK, Relton CL, Gaunt TR, Richardson TG. 2020. Characterizing the causal pathway for genetic variants associated with neurological phenotypes using human brain-derived proteome data. *Am J Hum Genet* **106:** 885–892. doi:10.1016/j.ajhg.2020.04.007

Kichaev G, Yang WY, Lindstrom S, Hormozdiari F, Eskin E, Price AL, Kraft P, Pasaniuc B. 2014. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet* **10:** e1004722. doi:10.1371/journal.pgen.1004722

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46:** 310–315. doi:10.1038/ng.2892

Koscielny G, An P, Carvalho-Silva D, Cham JA, Fumis L, Gasparyan R, Hasan S, Karamanis N, Maguire M, Papa E, et al. 2017. Open targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res* **45:** D985–D994. doi:10.1093/nar/gkw1055

Landrum MJ, Kattman BL. 2018. ClinVar at five years: delivering on the promise. *Hum Mutat* **39:** 1623–1630. doi:10.1002/humu.23641

Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR. 2012. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28:** 2540–2542. doi:10.1093/bioinformatics/bts474

Manolio TA. 2010. Genomewide association studies and assessment of the risk of disease. *N Engl J Med* **363:** 166–176. doi:10.1056/NEJMra0905980

McGowan LM, Smith GD, Gaunt TR, Richardson TG. 2019. Integrating Mendelian randomization and multiple-trait colocalization to uncover cell-specific inflammatory drivers of autoimmune and atopic disease. *Hum Mol Genet* **28:** 3293–3300. doi:10.1093/hmg/ddz155

Millwood IY, Walters RG, Mei XW, Guo Y, Yang L, Bian Z, Bennett DA, Chen Y, Dong C, Hu R, et al. 2019. Conventional and genetic evidence on alcohol and vascular disease aetiology: a prospective study of 500,000 men and women in China. *Lancet* **393:** 1831–1842. doi:10.1016/S0140-6736(18)31772-0

* Munafò MR. 2020. Triangulation of evidence. *Cold Spring Harb Perspect Med* doi: 10.1101/cshperspect.a040659

Munafò MR, Davey Smith G. 2018. Robust research needs many lines of evidence. *Nature* **553:** 399–401. doi:10.1038/d41586-018-01023-3

Ng PC, Henikoff S. 2003. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31:** 3812–3814. doi:10.1093/nar/gkg509

O'Connor LJ, Price AL. 2018. Distinguishing genetic correlation from causation across 52 diseases and complex traits. *Nat Genet* **50:** 1728–1734. doi:10.1038/s41588-018-0255-0

Oliphant A, Barker DL, Stuelpnagel JR, Chee MS. 2002. BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. *BioTechniques* **32:** 56–61.

Pingault JB, O'Reilly PF, Schoeler T, Ploubidis GB, Rijsdijk F, Dudbridge F. 2018. Using genetic data to strengthen causal inference in observational research. *Nat Rev Genet* **19:** 566–580. doi:10.1038/s41576-018-0020-3

* Porcu E, Sjaarda J, Lepik K, Carmeli C, Darrous L, Sulc J, Mounier N, Kutalik Z. 2020. Causal inference methods to integrate omics and complex traits. *Cold Spring Harb Perspect Med* doi:10.1101/cshperspect.a040493

Richardson TG, Harrison S, Hemani G, Davey Smith G. 2019a. An atlas of polygenic risk score associations to highlight putative causal relationships across the human phenome. *eLife* **8:** e43657. doi:10.7554/eLife.43657

Richardson TG, Richmond RC, North TL, Hemani G, Davey Smith G, Sharp GC, Relton CL. 2019b. An integrative approach to detect epigenetic mechanisms that putatively mediate the influence of lifestyle exposures on disease susceptibility. *Int J Epidemiol* **48:** 887–898. doi:10.1093/ije/dyz119

Richardson TG, Hemani G, Gaunt TR, Relton CL, Davey Smith G. 2020. A transcriptome-wide Mendelian randomization study to uncover tissue-dependent regulatory mechanisms across the human phenome. *Nat Commun* **11:** 185. doi:10.1038/s41467-019-13921-9

Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GL, Edwards KJ, Day IN, Gaunt TR. 2013. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat* **34:** 57–65. doi:10.1002/humu.22225

Speed D, Cai N; the UCLEB Consortium; Johnson MR, Nejentsev S, Balding DJ. 2017. Reevaluation of SNP heritability in complex human traits. *Nat Genet* **49:** 986–992. doi:10.1038/ng.3865

Staley JR, Blackshaw J, Kamat MA, Ellis S, Surendran P, Sun BB, Paul DS, Freitag D, Burgess S, Danesh J, et al. 2016. PhenoScanner: a database of human genotype-phenotype associations. *Bioinformatics* **32:** 3207–3209. doi:10.1093/bioinformatics/btw373

Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, Stein TI, Nudel R, Lieder I, Mazor Y, et al. 2016. The GeneCards Suite: from gene data mining to disease genome sequence analyses. *Curr Protoc Bioinformatics* **54:** 1.30.1–1.30.33. doi:10.1002/cpbi.5

Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, et al. 2015. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12:** e1001779. doi:10.1371/journal.pmed.1001779

Syvänen AC. 2005. Toward genome-wide SNP genotyping. *Nat Genet* **37:** S5–S10. doi:10.1038/ng1558

Taylor AE, Morris RW, Fluharty ME, Bjorngaard JH, Åsvold BO, Gabrielsen ME, Campbell A, Marioni R, Kumari M, Hällfors J, et al. 2014. Stratification by smoking status reveals an association of CHRNA5-A3-B4 genotype with body mass index in never smokers. *PLoS Genet* **10:** e1004799. doi:10.1371/journal.pgen.1004799

Taylor K, Davey Smith G, Relton CL, Gaunt TR, Richardson TG. 2019. Prioritizing putative influential genes in cardiovascular disease susceptibility by applying tissue-specific Mendelian randomization. *Genome Med* **11:** 6. doi:10.1186/s13073-019-0613-2

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491:** 56–65. doi:10.1038/nature11632

Trenkmann M. 2018. Lessons from 1 million genomes. *Nat Rev Genet* **19:** 592–593. doi:10.1038/s41576-018-0047-5

Verbanck M, Chen CY, Neale B, Do R. 2018. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat Genet* **50:** 693–698. doi:10.1038/s41588-018-0099-7

Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 2017. 10 years of GWAS discovery: biology, function, and translation. *Am J Hum Genet* **101:** 5–22. doi:10.1016/j.ajhg.2017.06.005

Wallace C. 2020. Eliciting priors and relaxing the single causal variant assumption in colocalization analyses. *PLoS Genet* **16:** e1008720. doi:10.1371/journal.pgen.1008720

Warrington NM, Freathy RM, Neale MC, Evans DM. 2018. Using structural equation modelling to jointly estimate maternal and fetal effects on birthweight in the UK Biobank. *Int J Epidemiol* **47:** 1229–1241. doi:10.1093/ije/dyy015

Watanabe K, Stringer S, Frei O, Umićević Mirkov M, De Leeuw C, Polderman TJC, Van Der Sluis S, Andreassen OA, Neale BM, Posthuma D. 2019. A global overview of

Cold Spring Harbor Perspectives in Medicine
www.perspectivesinmedicine.org

pleiotropy and genetic architecture in complex traits. *Nat Genet* **51:** 1339–1348. doi:10.1038/s41588-019-0481-0

Wen X, Pique-Regi R, Luca F. 2017. Integrating molecular QTL data into genome-wide genetic association analysis: probabilistic assessment of enrichment and colocalization. *PLoS Genet* **13:** e1006646. doi:10.1371/journal.pgen.1006646

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, et al. 2010. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42:** 565–569. doi:10.1038/ng.608

Yavorska OO, Burgess S. 2017. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *Int J Epidemiol* **46:** 1734–1739. doi:10.1093/ije/dyx034

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, et al. 2018. Ensembl 2018. *Nucleic Acids Res* **46:** D754–D761. doi:10.1093/nar/gkx1098

Zhao Q, Wang J, Hemani G, Bowden J, Small DS. 2019. Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. arXiv 1801.09652

Zheng J, Erzurumluoglu AM, Elsworth BL, Kemp JP, Howe L, Haycock PC, Hemani G, Tansey K, Laurin C; Early Genetics and Lifecourse Epidemiology (EAGLE) Eczema Consortium; et al. 2017. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33:** 272–279. doi:10.1093/bioinformatics/btw613

Zheng J, Haberland V, Baird D, Walker V, Haycock PC, Hurle MR, Gutteridge A, Erols P, Liu Y, Lou S, et al. 2020. Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nat Genet* **52:** 1122–1131.

Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, Montgomery GW, Goddard ME, Wray NR, Visscher PM, et al. 2016. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* **48:** 481–487. doi:10.1038/ng.3538

Zhu Z, Zheng Z, Zhang F, Wu Y, Trzaskowski M, Maier R, Robinson MR, McGrath JJ, Visscher PM, Wray NR, et al. 2018. Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat Commun* **9:** 224. doi:10.1038/s41467-017-02317-2

Cold Spring Harbor Perspectives in Medicine
www.perspectivesinmedicine.org