Jan 27, 2026

Human Accelerated Regions (HARs)-aware causal inference: Two-sample Mendelian randomization & colocalization for microbiome/metabolites

DOI
https://dx.doi.org/10.17504/protocols.io.x54v9bdqpl3e/v1

Siddharth Singh[1]

[1]Indian Institute of Technology Indore

Siddharth Singh
Indian Institute of Technology Indore

DOI: https://dx.doi.org/10.17504/protocols.io.x54v9bdqpl3e/v1

Protocol Citation: Siddharth Singh 2026. Human Accelerated Regions (HARs)-aware causal inference: Two-sample Mendelian randomization & colocalization for microbiome/metabolites . **protocols.io** https://dx.doi.org/10.17504/protocols.io.x54v9bdqpl3e/v1

Protocol status: In development
We are still developing and optimizing this protocol

Created: January 11, 2026

Last Modified: January 27, 2026

Protocol Integer ID: 238422

Keywords: ancestry harmonization before analysis, microbiome trait, exposure gwa, ancestry harmonization, microbiome, explicit harmonization exclusion report, microbial functional feature, host genetic liability, regional summary data, colocalization for microbiome, circulating metabolite, genetic liability, data provenance, using twosamplemr allele alignment rule, twosamplemr allele alignment rule, sample mendelian randomization

Abstract

This protocol implements a conservative, end-to-end workflow to test whether host genetic liability to gut microbiome traits, microbial functional features, or circulating metabolites causally influences neurodevelopmental and psychiatric phenotypes using two-sample Mendelian randomization (MR) with mandatory colocalization for headline claims. It standardizes data provenance via a manifest and enforces build and ancestry harmonization before analysis. Instruments are selected from exposure GWAS, LD clumped using TwoSampleMR::clump_data (PLINK-style clumping, with optional local reference), and harmonized to outcome summary statistics using TwoSampleMR allele alignment rules, producing an explicit harmonization exclusion report. Outcomes and regional summary data can be retrieved reproducibly from the OpenGWAS infrastructure using ieugwasr with access dates recorded. Primary MR uses IVW with a minimum sensitivity set (weighted median, MR-Egger, heterogeneity and pleiotropy diagnostics, directionality checks). Colocalization is then performed per locus using Bayesian ABF (coloc.abf, single causal variant) and escalated to SuSiE-based colocalization when multiple signals are plausible, requiring an LD matrix. A HAR-aware module overlays colocalized credible sets onto HAR interval annotations to prioritize loci with plausible neurodevelopmental regulatory relevance, explicitly treating HAR overlap as a heuristic rather than causal proof. Reporting outputs map directly to STROBE-MR items to support peer review defense and reuse.

Troubleshooting

Problem

No genome-wide significant instruments for exposure

Solution

It's common with microbiome GWAS that few or no SNPs reach 5×10^{-8}. If this happens, you have a few options: (1) Use a suggestive threshold (e.g., 1×10^{-5}) to obtain instruments, but be cautious, perform extensive sensitivity (the false positive rate is higher). (2) Increase the sample size by meta-analyzing additional cohorts if available, or (3) acknowledge that this exposure cannot be reliably tested with MR due to a lack of strong instruments. If using suggestive instruments, check their F statistics; if many are <10, results may suffer weak instrument bias (tending towards null); consider using methods like SIMEX extrapolation (not covered) or just interpret carefully as exploratory.

Problem

Instrument SNP not found in outcome data

Solution

This can happen if the GWAS outcome uses a more stringent SNP inclusion criterion or a different array. If an instrument is missing, try finding a proxy SNP in high LD. Use LDlink (web or API) or a reference panel, e.g., identify a proxy with r^2 > 0.8 in the European 1000G. Replace the missing SNP with the proxy SNP (and update the beta and alleles accordingly in the exposure data). If no good proxy, drop the SNP from MR and note the reduction in instrument count. If many instruments are missing, perhaps the outcome data are lower-density (e.g., only genotyped SNPs). In such a case, try to get the imputed/full outcome summary stats if possible.

Problem

Genome build mismatches leading to allele issues

Solution

If you notice weird allele alignments (like A vs C when it should be A vs T, etc.), double-check the genome build alignment. Perhaps liftover didn't map some SNPs correctly, or the rsIDs changed. Cross-verify a few sentinel SNP positions via dbSNP or UCSC. If necessary, convert the outcome coordinates instead, or use rsID joining rather than positions. Always ensure the allele pairing is correct before MR (garbage in, garbage out).

Problem

Heterogeneity and pleiotropy

Solution

If Cochran's Q is significant or MR-Egger intercept p < 0.05, this suggests pleiotropy (instruments affecting outcome via other pathways). To troubleshoot: examine which SNPs contribute to heterogeneity, use leave-one-out as described. If one SNP, when removed, makes Q disappear and significantly alters the effect, that

SNP is likely pleiotropic or an outlier. In an extended pipeline, use MR-PRESSO to attempt to remove outliers. If, after removing outliers, the causal effect remains, you can be more confident. If the effect disappears, the original MR result was driven by a problematic SNP. You should report that and perhaps not consider that a robust causal finding. Also, check phenotypes of instruments via PhenoScanner or GWAS Catalog; if an instrument SNP is known to affect a trait closely related to the outcome (other than via exposure), that's a red flag (e.g., a BMI gene used as an instrument for microbiome could affect outcome via BMI, not microbiome). Consider excluding such SNPs if justified, or use multivariable MR, including that factor if data permits.

Problem

Colocalization yields low PP.H4 despite significant MR

Solution

This is a critical scenario. It could mean the MR result is a false positive or that horizontal pleiotropy is at play (MR saw an association, but the underlying variants differ). If this occurs, first check whether the exposure and outcome association signals are indeed distinct: examine their lead SNPs and their correlation. Perhaps the exposure's lead SNP is modest, and the outcome's lead is another SNP in LD. If the r^2 between them is low, that explains low colocalization. In such a case, MR may be picking up a correlated signal. You might try a narrower MR (single SNP if one looks clearly the candidate) or just conclude that no solid evidence of a shared causal variant means the causal inference is not strongly supported. It's okay to have a significant MR that doesn't colocalize; explain it as likely confounded by linkage. If you suspect multiple causal variants, use the SuSiE approach: it might reveal a hidden colocalized component that simple coloc missed. If still nothing, then downgrade the causality claim.

Problem

SuSiE fine-mapping issues

Solution

SuSiE can fail to converge or give warnings, especially if the LD matrix is singular or the data has issues. To troubleshoot, ensure you provide a well-conditioned LD matrix (exclude variants in perfect LD or use a smaller region). You can also increase max_iter or set residual_variance if known. If SuSiE results in fewer credible sets than expected (like it only finds one when you suspect two), it might have absorbed multiple signals into one; you could try forcing more components (the L parameter). But don't over-fit. If it's too problematic, stick to coloc.abf and explain multi-signal possibility qualitatively.

Problem

HAR annotation pitfalls

Solution

HAR coordinates might be based on hg19; ensure you lifted them if your data is hg38, otherwise overlaps will be wrong. If using a broad overlap (± some kb), decide a threshold and stick to it (we used direct overlap or possibly ±20kb). If too many loci are near HARs by chance (HARs are ~0.023% of the genome, so chance

overlap is low but possible), prioritize those within HAR boundaries. If no loci hit a HAR, that's fine; not every result involves HARs. Don't force a connection; just report that none was found in that case.

Problem

Large data handling

Solution

If any GWAS file is extremely large (e.g., millions of SNPs), reading into R may be slow. Consider using tools to pre-filter by region (e.g., tabix if the file is indexed: tabix -h bigGWAS.vcf.gz chr1:1230000-1780000 > region.vcf to get a region for colocalization). Or use the BigQuery approach if supported by the data source. Memory errors in R can be mitigated by using data.table streaming or processing in chunks. For MR, since we ultimately only need significant SNPs from exposure and corresponding ones from outcome, it's feasible. For colocalization, restrict to regions of interest rather than the whole genome.

Problem

Results interpretation and presentation

Solution

If results seem too good to be true (e.g., extremely low p-values for MR with very few SNPs), double-check for errors. Extremely significant MR results for microbiome traits may indicate confounding (e.g., population stratification not corrected in GWAS or overlap). Always verify that the exposure GWAS had proper controls (e.g., principal components for ancestry). If not, population structure could produce spurious correlations (e.g., between gene frequencies and differences in the microbiome by ethnicity). In such a case, be extra careful; consider using MR Egger (it can sometimes detect unbalanced pleiotropy that might come from stratification).

Problem

Reproducibility issues

Solution

If rerunning yields slightly different results (especially in SuSiE), check that you set the random seed before those steps. SusieR's algorithm may have randomness; set the seed to get consistent credible sets. If using parallel threads, the order of results might change; ensure you collect results in a deterministic way.

Problem

Software updates

Solution

If you update R packages, results (especially colocalization or MR-PRESSO) might change due to algorithm improvements. That's why saving sessionInfo is crucial. If needed, stick to a known package version (e.g., install a specific version) for reproducibility across collaborators.

Before start

AppendixA. STROBE-MR Checklist (summary)

When reporting results from this MR analysis, adhere to the STROBE-MR guidelines. Key checklist items include:

- **Title/Abstract:** Mention Mendelian Randomization as the study design in the title and abstract, including the primary purpose (e.g., testing causal effect of microbiome on trait).
- **Background (Intro):** Explain scientific background and rationale, describe the exposure and why a causal relationship with outcome is plausible, and justify MR as an appropriate method.
- **Objectives:** State specific hypotheses and that MR estimates causal effects (prespecified directions).
- **Methods - Data Sources/Participants:** Describe study design (two-sample MR), populations of GWAS (sample sizes, cohorts, ancestry), and eligibility or selection criteria for variants (p-value thresholds, etc.). Detail measurement and quality control of genetic variants (clumping, palindromic SNP handling) and data harmonization.
- **Methods - Statistical Analysis:** Explicitly state the 3 core IV assumptions (relevance, independence, exclusion restriction). Describe how SNP-exposure and SNP-outcome data were harmonized and the MR methods used (IVW, Egger, etc.). Mention how multiple testing was addressed (e.g., FDR). List software and packages with versions.
- **Assessment of assumptions:** Note any methods used to examine IV assumptions, e.g., checking pleiotropy via Egger intercept or heterogeneity (Q test).
- **Sensitivity analyses:** Report what sensitivity or additional analyses were done (e.g., leave-one-out, reverse MR, colocalization, subgroup analyses).
- **Results - Descriptive:** Provide summary stats of exposure and outcome data (sample sizes, variance explained by instruments). If 2-sample MR, justify that the exposure and outcome samples are comparable or any overlap (ideally none).
- **Results - Main findings:** Report association of genetic instruments with exposure and outcome (e.g., first stage F stats, R^2). Then report MR causal estimates (beta/OR, CI) for exposure on outcome. Include measures of uncertainty and units (e.g., OR per SD of exposure). If helpful, translate effect into absolute risk difference (if outcome is disease). Provide plots (scatter, forest) as needed for illustration. Report any additional stats like heterogeneity Q.
- **Results - Sensitivity:** Summarize results of sensitivity analyses: e.g., “Egger intercept p=0.4 (no evidence of pleiotropy); leave-one-out showed no single SNP driving effect; reverse MR was null”. Mention if any results differed under robust methods (e.g., median vs IVW). If any causal effect was consistent with RCT or other evidence, note that.
- **Key results:** Summarize the key causal findings in a sentence or two relating back to objectives.
- **Limitations:** Discuss limitations, including potential violations of MR assumptions, pleiotropy, weak instruments, sample bias, etc. and general caution that association ≠ definitive proof.
- **Interpretation:** Give a cautious interpretation in context of other evidence. Discuss biological mechanisms that could underlie the causal link, if known, and whether the gene-environment equivalence assumption holds (for microbes, this is tricky; discuss how genetic proxies relate to actual microbial exposure). Use causal language carefully, acknowledging IV assumptions.
- **Generalisability:** Note if results likely generalize to other populations or are specific to the population studied.
- **Funding/Data:** Disclose funding sources and data availability (where can others get the GWAS data used, etc.).
- **Conflicts of interest:** Declare any, if applicable.

(For full details, see STROBE-MR Statement and Explanation documents. Ensuring these items are covered will increase transparency and trust in your MR findings.)

B. Default parameters used in this workflow

- **Genome build:** GRCh37 (hg19) assumed for input summary stats, unless specified. Liftover performed to harmonize builds if needed.
- **Significance threshold for instruments:** p < 5 × 10^{-8} (genome-wide significance). Extended pipeline allows p < 1 × 10^{-5} for suggestive instruments if GWS not available.
- **LD clumping:** r^2 < 0.001 within 1,000 kb window (approximately independent genome-wide) for instrument selection.
- **Minimum minor allele frequency:** 1% (SNPs with MAF < 0.01 removed to avoid rare variant issues).
- **MR methods:** IVW (random-effects) as primary; MR-Egger, weighted median as secondary; simple median and mode optional.
- **Heterogeneity test:** Cochran's Q with p<0.05 considered significant heterogeneity.
- **Pleiotropy test:** MR-Egger intercept p<0.05 considered evidence of directional pleiotropy.
- **Steiger directionality:** Use p<0.05 to flag potential reverse causation (or use the provided logical output from TwoSampleMR).
- **Colocalization window:** ±500,000 bp around lead SNP (adjustable per locus if needed).
- **Colocalization priors:** p1 = 1×10^{-4}, p2 = 1×10^{-4} (prior probabilities a given variant is associated with exposure or outcome, respectively), p12 = 1×10^{-5} (prior that variant is associated with both). These are slightly conservative defaults; can adjust if traits have known polygenicity.
- **Coloc PP threshold:** PP.H4 > 0.8 considered high evidence of colocalization; 0.5-0.8 moderate; <0.5 low.
- **Fine-mapping (SuSiE):** Maximum of 5 causal components tested by default (L=5); credible set coverage 95%. LD reference: 1000G Phase3 EUR.
- **HAR overlap distance:** 0 bp (require SNP to fall inside HAR region to call overlap). HAR regions typically as defined in input file (often a few hundred bp each).
- **Multiple testing correction:** FDR (Benjamini-Hochberg) across all exposure-outcome tests, q<0.05 denotes significant causal finding.
- **Plots:** Generate MR scatter and leave-one-out plots for any result with p<0.05; generate coloc regional plot for each tested locus.
- **Random seed:** 1234 for reproducible random operations (particularly for SuSiE fine-mapping).
- **Logging:** Verbose mode on for all major functions; write log messages to console and to logs/ files.

(Users may modify these defaults in the config.yaml *or script setup for different scenarios. If a more stringent instrument threshold is desired (p<1e-9) to reduce weak instruments, or a larger window for coloc if LD extends beyond 500 kb in a region, etc.)*

C. Manifest template snippets

manifest_inputs.csv (CSV with columns: *Dataset_ID*, *Description*, *Type*, *Population*, *N*, *File_Path*, *Source*):
(Note: N is sample size; for case-control, specify cases/controls if needed. Source gives citation or URL)
Each row corresponds to an input used. This manifest can be read by the script to automatically load data and assign to variables.
Optional: separate manifests for exposures and outcomes could also be used if looping through many of each.

D. Example file tree structure

After running the protocol, the project directory might be organized as follows:

```
HAR_causal_MR_Project/
├── config/
│   ├── config.yaml              # Configuration of parameters and file paths
│   └── manifest_inputs.csv      # Manifest of all input datasets (exposures,
outcomes, annotations)
├── data/
│   ├── exposures/
│   │   ├── MiBioGen_Bifidobacterium.txt         # Exposure GWAS summary stats
(example)
│   │   └── FHS_GABA.txt                         # Another exposure (metabolite) GWAS
│   ├── outcomes/
│   │   ├── PGC2_SCZ.sumstats.gz                 # Outcome GWAS summary
(schizophrenia)
│   │   └── ADHD2022.sumstats.txt                # Outcome GWAS summary (ADHD)
│   └── annotations/
│       ├── HAR_hg19.bed                         # HAR regions, hg19 coordinates
│       ├── genes_gencode.v19.bed               # Gene annotation (if used for
mapping)
│       └── Brain_eQTL_topHits.txt               # Brain eQTL significant SNPs
├── analysis/
│   ├── MR_workflow.Rmd             # The R Markdown or script containing this
protocol's code
│   └── MR_workflow.html            # (If using RMarkdown, the knitted HTML report)
├── results/
│   ├── MR/
│   │   ├── MR_results_all.csv           # Summary of MR estimates for all tested
pairs
│   │   ├── Bifidobacterium_SCZ_scatter.png    # MR scatter plot for a specific
exposure-outcome
│   │   ├── Bifidobacterium_SCZ_leaveoneout.png # Leave-one-out plot
│   │   └── ... (other plots/tables per pair)
│   ├── coloc/
│   │   ├── Bifidobacterium_SCZ_chr5_coloc.txt   # Coloc result for locus on chr5
│   │   ├── Bifidobacterium_SCZ_chr5_plot.png    # Regional association plot for that
locus
│   │   ├── ... (other loci and pairs)
│   └── annotations/
│       ├── prioritized_hits_table.csv    # Table of prioritized loci with HAR & eQTL
annotations
│       └── locus_brainEQTL_details.txt   # Any detailed eQTL findings per locus
└── logs/
    ├── MR_analysis.log           # Log file capturing console output (if saved)
    └── sessionInfo.txt           # R session info with package
versions:contentReference[oaicite:114]{index=114}
```

Description

1 Human Accelerated Regions (HARs) are rapidly evolving human genomic loci implicated in brain development and human-specific traits. Recent studies suggest certain HAR-linked genetic factors may influence gut microbiota composition, hinting at co-evolution between the human genome and commensal microbes. This computational protocol provides a step-by-step workflow for assessing causal relationships between microbiome or metabolite exposures and neurodevelopmental and psychiatric outcomes, while integrating HAR annotations to provide evolutionary context. We employ a two-sample Mendelian randomization (MR) approach, using genetic variants as instruments for exposures, coupled with colocalization analysis to ensure that any causal signal is driven by shared variants rather than linkage or pleiotropy. The pipeline is **modular and reproducible**, enabling adaptation to different exposure–outcome pairs. We outline procedures for obtaining GWAS summary statistics of microbial taxa (e.g., MiBioGen consortium data), microbial functional pathways, and metabolites (e.g., AGORA2 resource or blood metabolite GWAS), alongside outcome GWAS for neurodevelopmental/psychiatric traits (e.g., autism, schizophrenia, depression). Ancestry and genome-build harmonization steps are included (supporting GRCh37, GRCh38, etc. via liftOver), ensuring proper alignment of variants across datasets.

We detail the instrument SNP selection, data harmonization (aligning effect alleles across exposure/outcome, removing ambiguous variants), and MR analysis using inverse-variance weighted (IVW) and robust methods (Egger, weighted median), and sensitivity tests (heterogeneity, MR-Egger intercept for pleiotropy, Steiger directionality test). We then describe Bayesian colocalization using both the Coloc approximate Bayes factor method and SuSiE fine-mapping to confirm whether the exposure and outcome share the same causal variant(s). A HAR-aware prioritization module overlays known HAR coordinates onto the loci of interest, maps putative causal variants/genes to HARs, and cross-references brain eQTL data to highlight functional relevance. Multiple-testing correction (false discovery rate, FDR 5%) is applied to any analyses involving numerous traits. Throughout, clear quality control (QC) gates and troubleshooting notes are provided (e.g., handling weak instruments, allele frequency mismatches, ancestral LD differences). The protocol distinguishes between a minimum-credible pipeline, focusing on essential MR and colocalization steps for a quick causal inference check, and an extended pipeline with comprehensive analyses (fine-mapping, extensive annotations, etc.) for in-depth exploration.

2 **Guidelines**

2.1 Study Design & Assumptions: This protocol implements a two-sample MR design, requiring that the exposure and outcome GWAS come from non-overlapping samples to

avoid bias. Verify that no individuals are shared between the exposure GWAS (e.g., microbiome study) and outcome GWAS (neuro/psych trait study). The MR analysis relies on three core assumptions: (1) genetic instruments are strongly associated with the exposure of interest; (2) instruments are not associated with confounders of the exposure-outcome relationship; (3) instruments affect the outcome only through the exposure (no horizontal pleiotropy). These assumptions should be explicitly considered and later assessed using sensitivity tests (e.g., MR-Egger intercept for pleiotropy). Use Steiger's directionality test to ensure the inferred causal direction is exposure & outcome.

2.2 Data Requirements: Obtain high-quality GWAS summary statistics for both exposures and outcomes. Exposures can include: (a) Microbiome taxa abundance, e.g. genus-level or species-level GWAS from MiBioGen consortium (the largest to date with 18,340 individuals across 212 taxa) or other cohorts; (b) Microbial functional pathways or gene modules, e.g. pathway summary scores or gene-based microbiome GWAS if available, or synthetic traits derived from resources like gutMGene (which links microbes, their metabolites, and host target genes); (c) Metabolites, e.g. circulating metabolite levels from metabolomics GWAS (such as those in blood, possibly related to microbiota e.g. short-chain fatty acids, neurotransmitters like GABA, etc.). Outcomes should be well-powered GWAS for neurodevelopmental or psychiatric traits (e.g., autism spectrum disorder, schizophrenia, bipolar disorder, depression, ADHD, cognitive performance, brain imaging phenotypes), ideally from large consortia (e.g., PGC, UK Biobank) to ensure sufficient instrument-outcome associations. Ensure each dataset includes at minimum: SNP identifier (rsID or chromosome:position), effect allele and non-effect allele, effect size (beta or OR) with standard error, p-value, and allele frequency.

2.3 Ancestry and Genome Build Harmonization: Check the ancestry of each GWAS. If exposure and outcome differ in ancestry (e.g., one European, one East Asian), you may analyze each ancestry separately or be cautious when interpreting cross-ancestry MR due to differences in linkage disequilibrium (LD) patterns. Ideally, use GWAS with comparable ancestry or perform a multi-ancestry analysis as an extension. Also, confirm the genome build of each summary statistic dataset (e.g., GRCh37/hg19 vs GRCh38/hg38). Use UCSC LiftOver or equivalent tools to convert variant coordinates so that exposure and outcome data use a consistent reference build before harmonization. Maintain backup copies of the original datasets and record any liftover procedures in the manifest (including the chain file used).

2.4 Instrument Selection: The default instrument selection criterion is genome-wide significance ($p < 5\times10^{-8}$) for association with the exposure trait. If no SNPs reach this threshold for a given exposure (common for microbiome traits with modest GWAS power), consider a secondary locus-wide significance threshold (e.g. $p<1\times10^{-5}$) for suggestive instruments, but note: using more lenient thresholds increases risk of weak instruments and false positives, so apply such only in an exploratory or extended analysis with appropriate caution. After p-value filtering, perform LD clumping to ensure instruments are approximately independent: e.g., remove SNPs in high LD ($r^2 > 0.001$

within a ±250 kb window), keeping the most significant SNP per locus. (Use PLINK or an MR package function for clumping, with a 1000 Genomes reference panel of matching ancestry for LD estimates.) Each exposure should ideally have multiple independent instruments; if only one instrument is available, MR can still be performed (using a Wald ratio or simple method), but results are less robust. Exclude instruments with extremely low minor allele frequency (MAF) (<1%) if not already filtered out in the source GWAS, to avoid unstable effect estimates.

2.5 Data Harmonization: Harmonize exposure and outcome summary statistics so that effect alleles and their effects correspond. This is critical to avoid strand mismatches or sign flips. Use an automated harmonization function (e.g., TwoSampleMR::harmonise_data in R) or a manual procedure: Align the datasets by SNP rsID. For any instrument SNP absent in the outcome dataset, attempt to find a suitable proxy in high LD (e.g., r^2 > 0.8) using an LD reference panel or a service like LDlink; record any proxy substitutions in the manifest. Ensure alleles match on the strand. If using rsIDs and both are aligned to the same reference genome, the harmonization function typically handles strand flips (e.g., G•C or A•T) and allele order. Remove palindromic SNPs (A/T or G/C with ~50% allele frequency) where strand ambiguity can't be resolved, or use allele frequency information to infer alignment if available (many MR tools do this internally). Confirm that for each instrument, the effect allele frequency is reasonably consistent between exposure and outcome datasets (discrepancies might indicate strand issues or mismatched build if liftover failed). By the end of harmonization, each instrument SNP will have: an effect allele, a beta (and SE) for exposure, and a beta (SE) for outcome corresponding to the same effect allele. The resulting harmonized dataset serves as the input to MR.

2.6 Parameter Configuration: Prepare a configuration file (e.g., config.yaml) or script section to set key analysis parameters. For reproducibility, explicitly document parameters such as significance thresholds (e.g. 5×10^{-8}), clumping r^2 cutoff (e.g. 0.001), clumping window (e.g. 1 Mb), colocalization window size (e.g. ±500 kb around index SNP), colocalization priors (default p1 = 1×10^{-4}, p2 = 1×10^{-4} for one causal variant in each trait, and p12 = 1×10^{-5} for a shared causal variant), and FDR significance level (e.g. 0.05). Setting these in one place (e.g., in a YAML file or at the top of the script) makes it easy to adjust for different scenarios or when running a minimum vs an extended analysis. Also specify file paths for data inputs, output directories, and external resources (e.g., the path to the liftOver chain files and a reference panel for LD, if needed). Use set.seed(...) at the start of any procedures involving randomness (e.g., the SuSiE algorithm or MR-PRESSO outlier tests) to ensure reproducibility.

Materials

3 Computing Environment:

3.1 A computer with R (version 4.0+ recommended) and/or Python (3.8+) installed. The pipeline can be implemented primarily in R using packages for MR and colocalization. Ensure you have enough memory and storage for large GWAS summary files (tens of gigabytes if many traits are included). High-performance computing (cluster or cloud) is recommended for heavy data operations (especially fine-mapping).

3.2 Required R Packages:

- *TwoSampleMR* (for MR analysis, MR Steiger test, etc.), *MendelianRandomization* (for alternate MR methods if needed), *MRPRESSO* (optional, for pleiotropy outlier correction), *coloc* (v5.1+ for colocalization, supports SuSiE), *susieR* (for fine-mapping if using SuSiE within coloc), *ieugwasr* (for accessing OpenGWAS API, optional if pulling data programmatically), *LDlinkR* (optional, to find LD proxies if needed), *liftOver* or *rtracklayer* (for genome build conversion), *data.table* (for efficient data handling), *dplyr/tidyr* (for data manipulation), *ggplot2* (for plotting results, e.g. Manhattan plots, scatter plots, etc.), *qvalue* or *p.adjust* in base R for FDR.
- If using Python, packages such as *pandas*, *numpy*, *statsmodels*, *pyranges*, *pyBigWig* (for genome annotations), and MR-specific libraries (e.g., *MendelianRandomization* for Python or custom scripts) may be used. The example here assumes R for core analyses.

3.3 **Command-line Tools:** PLINK 1.9 or 2.0 (for clumping and LD checks if needed outside R), tabix (if working with indexed GWAS files), liftOver utility from UCSC (if not using R packages for that), bcftools (if working with GWAS in VCF format). Ensure these are in your PATH if you use them.

3.4 **Data Inputs:**
Exposure GWAS summary statistics: One or more files containing summary-level association results of host genetic variants with microbiome or metabolite traits:

- **Microbiome composition:** Obtain from the MiBioGen consortium (Kurilshikov et al., Nature Genetics 2021). This meta-analysis of ~18k individuals provides summary stats for 212 taxa (131 genera, plus higher taxa). Data can be downloaded from the MiBioGen website (e.g., mibiogen.org) or via the GWAS catalog/OpenGWAS (if available). Each trait's file is typically a text file with columns for SNP, alleles, effect, p, etc. Another source: the TwinsUK Registry microbiome GWAS (if interested in species-level, e.g., four species in ~1,126 twins). Ensure to get the correct genome build (MiBioGen is on GRCh37) and note ancestry (multi-ethnic, predominantly European).
- **Microbial functional traits:** If analyzing pathways or microbial gene content, you might use summary stats from a GWAS of functional pathway abundance. If not directly available, one strategy is to derive pathway quantitative traits per individual (e.g., via PICRUSt or shotgun metagenomic analysis) and then use a public GWAS (if done) or perform one if raw data is available (not covered here). Alternatively, use the gutMGene database (Yang et al., Nucleic Acids Res 2022) as an annotation resource to identify candidate microbe-related gene targets; while gutMGene is not a GWAS, it

can connect microbial metabolites and host genes. For this protocol, we assume you have, or will focus on, a particular microbe-associated function with known genetic instruments (e.g.,using known host loci that affect the production of a microbiota-dependent metabolite).

- **Metabolites:** If examining metabolites (e.g., short-chain fatty acids, neurotransmitters, etc.), obtain GWAS summary stats for those. For instance, the Framingham Heart Study (FHS) published GWAS of certain metabolites influenced by diet/microbiome (e.g, choline, carnitine, beta-hydroxybutyrate, GABA, serotonin, TMAO). Another resource: the AGORA2 metabolic models – though not a GWAS, AGORA2 (Magnusdottir et al., 2020) can help identify microbe contributions to metabolite pathways. For purely genetic analysis, prefer large metabolite GWAS (e.g., Shin et al., 2014, for serum metabolites, or newer studies in Nature or Nature Genetics for metabolomics QTLs). Ensure files are summary statistics with SNP-level data (not just genetic correlation results).

3.5 **Outcome GWAS summary statistics:** Summary data for neurodevelopmental or psychiatric phenotypes. Examples (to be retrieved as needed):

- **Neurodevelopmental disorders**: e.g., Autism Spectrum Disorder GWAS (e.g., Grove et al. 2019, or latest updates from the Autism Working Group), Attention-Deficit/Hyperactivity Disorder (ADHD) GWAS (Demontis et al. 2019 or newer), Intellectual disability or developmental delay if available.
- **Psychiatric disorders**: e.g., Schizophrenia (e.g., PGC 2022 mega-analysis), Bipolar Disorder, Major Depressive Disorder (e.g., 2019 Wray et al. GWAS or later), Anxiety, etc.
- **Other brain traits**: optionally, IQ/cognitive performance, educational attainment (as proxies for neurodevelopment), or brain imaging phenotypes (from UK Biobank) could be considered outcomes if relevant to the hypothesis.
- These data are often available via the Psychiatric Genomics Consortium (PGC) website or the GWAS Catalog. Many are also accessible through the **IEU OpenGWAS** project. OpenGWAS provides API access to numerous GWAS. You can use trait IDs (e.g., ieu-b-** or ukb-b-** IDs) to download VCF or summary files. In this workflow, either download the summary text files or use ieugwasr::gwas_open() to query needed SNPs.
- Make sure to note sample sizes and ancestry of outcome GWAS; if substantially different from exposure GWAS ancestry, stratify analyses or interpret with caution as noted.

3.6 Reference files and annotations:

- **HAR coordinates:** A BED file or equivalent listing the genomic coordinates of known Human Accelerated Regions. You can retrieve a curated list from the literature, e.g., Doan et al. (Cell, 2016) reported ~2,700 HARs. The Pollard lab (UCSC/Gladstone) or supplementary material of recent HAR studies (e.g., the 2023 Neuron paper by Pollard's team) may provide coordinates (often in hg19). Alternatively, the UCSC Genome Browser has a public track for HARs (e.g., Pollard HARs) which can be

downloaded via the Table Browser (specify output as BED). For this protocol, assume we have HAR_hg19.bed with the following columns: chr, start, end, and HAR_ID. If your GWAS are in hg38, use LiftOver to get HAR coordinates in hg38. These will be used to flag if any identified loci overlap HARs.

- **Gene annotation and brain eQTL data:** To map significant variants or colocalized signals to genes, have gene reference data (e.g., GENCODE or RefGene). For eQTLs, obtain summary data or significant eQTL lists for brain tissues, e.g., from GTEx v8 (multiple brain regions), PsychENCODE, or other brain eQTL databases. You might not need full eQTL summary stats; a common approach is to query if top SNPs are eQTLs for any gene in the brain. For simplicity, you can use a precompiled list of top brain eQTLs (if available) or query the GTEx Portal for candidate SNP-gene pairs. Another useful resource is OpenTargets or QTL Catalog for colocalization with eQTLs, but here we will assume manual checking of a few prioritized loci.
- **Liftover chain files:** If needed, download chain files from UCSC (e.g., hg19ToHg38.over.chain.gz) to perform coordinate conversion of GWAS or annotation files.
- **1000 Genomes LD reference:** If using LD-based methods (clumping, finding proxies, or running SuSiE fine-mapping), have access to a reference genotype panel in the relevant population (1000 Genomes Phase3, or UKB imputed data if allowed). 1000G EUR can be used for the European LD structure. PLINK .bed/.bim/.fam or VCF format works. If using the coloc-SuSiE approach, you may need allele correlation (LD) information for the region; some implementations (like coloc.susie in the coloc R package) accept the summary stats and infer LD via the susie_rss function given an LD matrix.

3.7 **Outputs / intermediate files:** Plan directories for results:

- results/MR/ for MR results (e.g., CSV files of causal estimates per exposure-outcome, plots of MR fit or funnel plots for each, etc.).
- results/coloc/ for colocalization outputs (posterior probabilities, credible sets, regional association plots).
- results/annotations/ for any HAR overlap or eQTL mapping results.
- logs/ for any log files and session info.
- These will be generated during the Steps; ensure you have write permissions.

4 Steps

4.1 Set up environment and load libraries:

Create a new R script or an R Markdown document for the analysis. Load required packages (see Materials):

```
library(TwoSampleMR)   # MR methods
library(coloc)         # Colocalization
library(susieR)        # Fine-mapping
library(data.table)    # Efficient data import
# ... (other libraries)
set.seed(1234)         # for reproducibility in any stochastic
steps
```

Ensure the working directory is set to your project folder. Read in configuration parameters (if using a config.yaml, load it via yaml::read_yaml("config.yaml")). Store default thresholds:

```
pval_threshold <- 5e-8
clump_r2 <- 0.001
clump_kb <- 1000
coloc_window <- 5e5    # 500kb window
coloc_p1 <- 1e-4; coloc_p2 <- 1e-4; coloc_p12 <- 1e-5
FDR_cutoff <- 0.05
```

Create output directories if not already present:

```
dir.create("results/MR", showWarnings = FALSE, recursive = TRUE)
dir.create("results/coloc", showWarnings = FALSE, recursive =
TRUE)
dir.create("results/annotations", showWarnings = FALSE, recursive
= TRUE)
dir.create("logs", showWarnings = FALSE)
```

Throughout the pipeline, check for any R warnings or errors and address them immediately

4.2 **Input data retrieval and formatting:**
Retrieve exposure and outcome GWAS summary statistics as per your manifest. If not already downloaded, use provided links or APIs. For example, to fetch from OpenGWAS:

```
# Example: Download schizophrenia GWAS (PGC ID ieu-b-1180) using
ieugwasr
library(ieugwasr)
scz_dat <- ieugwasr::gwas_open("ieu-b-1180")  # returns a
connection or data object
```

If files are local, use fread() from data.table for fast loading (many GWAS are large). For each exposure and outcome file, load them into R data frames or data.table. Standardize column names for ease: e.g., ensure we have SNP, chr, pos, effect_allele, other_allele, beta, se, p, eaf (effect allele freq) where possible.
Perform basic QC on loaded data: remove any variants with missing values, ensure all p-values are within [0,1], etc. For exposure traits, it may be useful to subset the data early to just genome-wide significant hits ($p < 5e-8$) to reduce memory usage, since others won't be used as instruments. However, keep a buffer around significance if you plan locus-wide selection ($p < 1e-5$ for suggestive).

For each exposure, record the total number of significant SNPs identified. For each outcome, note the sample size (some MR functions require it for certain analyses like Steiger test).

4.3 **Liftover and coordinate alignment (if needed):**
If your exposure and outcome data are on different genome builds, convert one to the other's coordinate system. It's often easiest to lift the smaller dataset to the build of the larger one. For example, if exposure GWAS is hg19 and outcome is hg38, lift exposure SNP positions to hg38:

```
library(rtracklayer)
chain <- import.chain("path/to/hg19ToHg38.over.chain")
GR_exposure <- GRanges(seqnames=Rle(exposure_dt$chr),
ranges=IRanges(exposure_dt$pos, exposure_dt$pos))
GR_exposure_lift <- liftOver(GR_exposure, chain)
```

This yields new coordinates for most SNPs. Merge the lifted positions back into the exposure data (some SNPs may not map; drop those or flag them). Alternatively, use rsIDs to match datasets if both use dbSNP IDs (this implicitly harmonizes build assuming rsIDs are consistent). Using rsIDs is recommended if available, since MR tools often use them to match variants.

After liftover or matching, ensure both datasets refer to identical reference alleles for

each SNP. If one dataset is from an older reference (e.g. non-reference alleles), aligning alleles in the next step will handle it.

4.4 **Instrument SNP selection for each exposure:**

For each exposure trait:

- **Identify candidate instruments:** Filter the exposure GWAS data to variants surpassing the significance threshold ($p < 5e-8$ by default). If none or very few SNPs meet this, consider the extended approach: use a more lenient threshold like $p < 1e-5$ to gather instruments, but be aware of weaker instruments. Record the threshold used.
- **LD Clumping:** On the filtered set, perform clumping to ensure independence. If using TwoSampleMR, you can use its built-in clump function which uses LD reference (1000 Genomes EUR by default):

```

```

where exposure_hits is a data frame of SNP, pval, etc. This returns a pruned set of SNPs. Alternatively, save SNP list and run PLINK:

```
plink --bfile 1000G_EUR --extract exposure_snps.txt --indep-
pairwise 1000 250 0.001 --out clump_out
```

and use the resulting .prune.in SNP list to filter.

- **Check instrument strength:** Compute or note F-statistics for each instrument as F = \frac{\beta^2}{\text{se}^2} (with 1 df) or use TwoSampleMR's add_rsq() to get variance explained and then F. Generally $F > 10$ indicates sufficient strength; if many instruments have $F < 10$, consider that exposure potentially underpowered and interpret MR with caution. (Weak instruments can bias MR towards null or cause inflation if many.)
- Save the list of selected instruments for each exposure to a file (e.g. results/MR/[exposure]_instruments.txt) for record-keeping. Include columns: SNP, effect allele, beta, se, p, etc. from exposure GWAS.
- If analyzing multiple exposures, repeat this step for each, and prepare to loop through exposures in subsequent steps. *Quality control:* If an exposure yields no instruments even at $p<1e-5$, you may drop that exposure from MR analysis (report as no instrumental evidence). If only 1-2 instruments, note that MR-Egger won't be applicable (needs ≥3 SNPs), and results will rely on IVW (or simple ratio if 1 SNP). Plan sensitivity accordingly.

4.5 **Harmonize exposure and outcome datasets (per exposure–outcome pair):**

Now, for each exposure-outcome combination, harmonize the data so that the effect of

the instrument on exposure and effect on outcome correspond to the same allele:

Use the function harmonise_data from TwoSampleMR for convenience: it takes a list of exposure instruments and the outcome summary data, and aligns alleles, accounting for strand flips and possible allele swaps:

```
exp_dat <- TwoSampleMR::format_data(exp_clumped, type="exposure",
snp_col="SNP", beta_col="beta", se_col="se",
effect_allele_col="effect_allele",
other_allele_col="other_allele", pval_col="p")
out_dat <- TwoSampleMR::format_data(outcome_data, type="outcome",
snp_col="SNP", ...)
harmonized <- TwoSampleMR::harmonise_data(exp_dat, out_dat)
```

This will remove palindromic SNPs with intermediate frequency by default and add a column indicating allele orientation.

- If not using TwoSampleMR, do manually: for each instrument SNP, find the matching SNP in outcome data (by SNP ID). If alleles don't match, flip the beta sign for outcome if the effect allele is the complementary base. For strand ambiguous SNPs (A/T or G/C), if minor allele frequency is available in both, infer alignment by comparing MAF (if one is ~1 – MAF of other, a flip is needed). When in doubt, exclude those SNPs.
- After harmonization, exclude any variant flagged for allele mismatch or duplicate. The final harmonized dataset should have, per SNP: beta_exposure, beta_outcome, se_exposure, se_outcome, effect_allele (aligned), other_allele, plus perhaps outcome's effect allele frequency if provided.
- *Sanity check:* Quickly inspect that no obviously impossible data remain (e.g., identical beta_exposure and beta_outcome values could indicate a mistake if traits are different, unless by chance). Also check the direction: if the exposure's increasing allele is associated with higher outcome (positive beta_outcome) or lower outcome (negative), this is the crude direction of effect.\
- Save the harmonized dataset (e.g., as .csv or RDS for each pair) if you plan to examine or reuse outside of R.

4.6 **Perform Mendelian Randomization analysis (IVW and additional methods):**
For each exposure–outcome pair, perform the MR tests:

- **Primary MR estimate (IVW):** Use inverse-variance weighted meta-analysis of SNP effects. In TwoSampleMR, this is done via mr(harmonized_data, method_list=c("mr_ivw")) or simply mr() which by default includes IVW. This yields an estimate of causal effect (e.g. in units of outcome per unit of exposure, depending on trait units) and a p-value. If the outcome is binary (disease), MR estimates are often

reported as odds ratio (which can be derived from the beta assuming a log-odds outcome). TwoSampleMR reports an OR if configured so; otherwise, you can exponentiate the beta for interpretation.

Secondary MR methods (robust to pleiotropy): It is good practice to run additional methods:

- MR-Egger regression: provides an estimate with a different assumption (allows non-zero intercept representing pleiotropy). Use method_list=c("mr_egger_regression"). Note that Egger often has larger uncertainty. Key output is also the intercept term (accessible via mr_pleiotropy_test(harmonized_data) in TwoSampleMR) – if the Egger intercept is significantly non-zero ($p < 0.05$), it suggests directional pleiotropy, meaning instruments collectively have pleiotropic effects on outcome violating MR assumptions.
- Weighted median: yields a consistent estimate even if up to 50% of weight comes from invalid instruments. Use mr_weighted_median.
- (Optional) Weighted mode or simple mode: other robust methods that can be tried (mr_weighted_mode).
- (Optional, extended) MR-PRESSO: perform the MR-PRESSO global test for pleiotropy and outlier correction if pleiotropy detected (requires the MRPRESSO package). This can identify if one or more SNPs are outliers contributing to heterogeneity, remove them, and re-estimate the causal effect.

Heterogeneity test: If multiple SNPs, examine Cochran's Q statistic for IVW (TwoSampleMR reports this via mr_heterogeneity(harmonized_data)). A significant Q ($p<0.05$) indicates heterogeneity in effect estimates beyond chance, could be due to pleiotropy or different causal mechanisms for different SNP subsets. High heterogeneity suggests caution in interpreting IVW; consider relying more on robust methods or pinpointing outliers.

- Collect results from these methods. You may tabulate: IVW beta, SE, p; Egger beta, SE, p and intercept p; Weighted median beta, etc. Also note number of SNPs (N_instruments). Store this in a results table (e.g., append to a data.frame with columns Exposure, Outcome, method, beta, OR (if binary outcome), 95% CI, p-value, N_snps, etc.).
- If running many pairs, automate the loop and save an intermediate CSV of all MR results for all trait pairs for record.
- *Thresholding:* After obtaining MR results, determine significance considering multiple testing. If only one exposure–outcome, use conventional $p<0.05$ (or 0.005 for more stringent). If many tests, apply FDR correction later (Step 12). In any case, record estimates and confidence intervals even for non-significant results, as they are informative for meta-analyses or comparisons.

```
res <- mr(harmonized, methods = c("mr_ivw", "mr_egger_regression",
"mr_weighted_median"))
res_single <- mr_singlesnp(harmonized)  # SNP-wise effects,
optional for funnel plot
pleio <- mr_pleiotropy_test(harmonized)
hetero <- mr_heterogeneity(harmonized)
```

These provide the needed outputs. Print or save them.

4.7 **Sensitivity analyses and interpretation checks:**
After obtaining MR estimates, perform additional checks for each significant or suggestive finding:

- **Directionality (Steiger test):** Use Steiger's approach to confirm the variance explained in exposure by the instruments is greater than the variance explained in outcome (if not, it suggests reverse causality). TwoSampleMR provides directionality_test() which returns a p-value or simply returns the logical conclusion (TRUE = correct direction):

```
steig <- directionality_test(harmonized)
```

If the test indicates the "outcome precedes exposure" (i.e., potential reverse causation), treat the causal claim with skepticism, it could be that the outcome actually influences the exposure (or there's residual confounding).

- **Single-SNP analyses:** Examine each instrument's individual Wald ratio (effect_on_outcome / effect_on_exposure) and its confidence interval. If one or two SNPs drive the entire MR signal (especially in heterogeneous result), note that. You might down-weight their influence or perform leave-one-out analysis:

```
loo <- mr_leaveoneout(harmonized)
```

This will compute IVW leaving each SNP out in turn. Plot the leave-one-out estimates to see if removing a particular SNP drastically changes the result, which would indicate that SNP is an outlier.

- **Plotting (Optional but recommended):** Generate an MR scatter plot: X-axis SNP effect on exposure, Y-axis SNP effect on outcome, with error bars. Add lines for IVW and Egger fits. This visualizes the overall trend and any outlier SNPs. Use TwoSampleMR::mr_scatter_plot or manually with ggplot. Also, a funnel plot of individual SNP effects can illustrate heterogeneity (available via mr_funnel_plot). Save plots in results/MR/ with informative names (e.g. scz_on_bifidobacterium_scatter.png).
- **Check alignments & units:** Ensure the effect estimates are interpretable: e.g., if exposure is binary or an increase per unit (like per SD of abundance). If needed, convert units so that a meaningful effect size is presented (like OR per fecal count

increase etc., though often just treating them as z-score units is fine). For metabolite exposures, consider if higher metabolite - > outcome effect is expected directionally or not. If something is counterintuitive, double-check if perhaps the effect allele is actually lowering the metabolite (sign issues) and thus interpretation might flip.

(At this point, you have MR results. If the minimum pipeline is being followed strictly, one could decide to stop at reporting MR results and note which exposure-outcome pairs show evidence of causality. However, to increase confidence, it's strongly advised to continue with colocalization to verify genetic alignment of signals.)

4.8 **Reverse causation analysis (bidirectional MR, optional):**

As an extended analysis, consider testing the reverse direction: could the outcome trait be causally influencing the microbiome or metabolite? This requires that the outcome trait has its own genetic instruments (which for polygenic diseases is true, there are many GWAS hits). If a robust set of instruments exists for the outcome, perform an MR treating outcome as "exposure" and microbiome feature as "outcome." If schizophrenia has genome-wide significant SNPs, test if genetically higher schizophrenia risk is associated with changes in gut microbiome abundance. Use the same steps above: select outcome instruments (p<5e-8 from outcome GWAS), harmonize with exposure (microbiome) GWAS, run MR. This helps identify if a nominal association might actually be due to outcome_exposure causality (reverse). If the reverse MR is also significant and of similar magnitude, it might indicate bidirectional or just unresolved confounding. Typically, one expects no effect in reverse if original direction was true (except in cases of feedback loops). Document any such tests and results. If none of the original MR results were significant, reverse MR is less priority, but still may be done for completeness (with FDR correction as well if multiple).

4.9 **Colocalization analysis (bayesian, single causal variant assumption):**
Next, perform colocalization to ensure that the exposure and outcome share the same causal variant in each genomic locus of interest. Focus on loci where MR found evidence of causality (or at least the exposure had significant instruments). For each locus (typically defined around each lead instrument SNP):

- **Define region:** e.g. a ±500 kb window around the lead SNP (instrument) or around the top SNP for outcome in that region. Ensure the region captures the full LD block of association. Use the coloc.abf method from the **coloc** R package. Gather all SNPs (and their summary stats) for both exposure and outcome within this window. If you have the full GWAS, simply filter by chromosome and position. If not, you may need to query a database or have pre-subset data. Both traits' data must be available for all SNPs in the region (or at least overlapping set).
- **Run coloc analysis:** Using coloc.abf(dataset1, dataset2), where dataset1 and dataset2 are lists with required info: $beta, $varbeta (variance of beta, i.e. SE^2), $N (sample size), $MAF (vector of minor allele freqs), $type ("quant" or "cc" for case-control with

prevalence if needed), and $snp (vector of SNP ids). TwoSampleMR has a helper coloc_abf() that might simplify this by extracting from loaded data frames. Set the prior probabilities; using the defaults p1=1e-4, p2=1e-4, p12=1e-5 is typical.

- **Interpret output:** The result will give posterior probabilities for hypotheses: H0 (no association either trait), H1 (assoc with trait1 only), H2 (trait2 only), H3 (both traits assoc but different SNPs), H4 (colocalized: both traits assoc and share causal SNP). Focus on **PP.H4,** the probability of colocalization. A common rule is PP.H4 > 0.8 (80%) is high evidence of shared causal variant. Also check that neither PP.H3 nor others are comparably high. If PP.H4 is low (e.g. < 0.5) while MR suggested an effect, this implies either multiple signals or a possible violation of MR assumptions (like horizontal pleiotropy). In such cases, the variant affecting exposure might not be the same driving outcome risk, so the causal interpretation is weakened.
- Repeat this colocalization for each locus corresponding to each exposure's instruments. If an exposure had multiple independent loci (instruments on different chromosomes, say), do each. If an instrument is in a huge LD block with multiple candidate signals, consider narrowing window or doing conditional analysis if data allows (out of scope for now).
- Save the colocalization results for each locus (for transparency). Also, record the top SNPs in the region for each trait, e.g. the SNP with highest association for outcome, to see if it matches the exposure's instrument SNP or is very close.

```
library(coloc)
exp_region <- exposure_data[exposure_data$chr == chr_i &
exposure_data$pos >= (pos_i-5e5) & exposure_data$pos <=
(pos_i+5e5), ]
out_region <- outcome_data[outcome_data$chr == chr_i &
outcome_data$pos >= (pos_i-5e5) & outcome_data$pos <= (pos_i+5e5),
]
coloc_res <- coloc.abf(
    dataset1=list(beta=exp_region$beta, varbeta=exp_region$se^2,
N=exp_sample_size, type="quant", snp=exp_region$SNP,
MAF=exp_region$maf),
    dataset2=list(beta=out_region$beta, varbeta=out_region$se^2,
N=out_sample_size, type="cc", snp=out_region$SNP,
MAF=out_region$maf),
    p1=1e-4, p2=1e-4, p12=1e-5
)
print(coloc_res$summary)  # contains PP for each hypothesis
```

- If PP.H4 >= 0.8, mark this locus as "colocalized". If in the gray zone (0.5-0.8), interpret as moderate evidence; consider further analysis with SuSiE (next step) or more data. If PP.H3 (two distinct signals) is high, conclude no colocalization,

exposure and outcome associations are likely driven by different variants despite being in the same region (could mean horizontal pleiotropy or just coincidence).

- For each locus, consider making a **regional plot** to visualize overlap: e.g. plot -log10p of exposure and outcome across the region, highlighting the lead SNP(s). This can be done with LocusZoom or manually with ggplot. If the peak positions overlap, that's visual support for colocalization (to complement the numeric PP). Save any such plots (e.g. results/coloc/[exposure]_[outcome]_chrX_coloc.png).

4.10 **Colocalization with fine-mapping (SuSiE for multiple causal variants, extended):**
In complex loci with possible multiple independent variants, use the SuSiE-based colocalization approach. This can identify colocalization even if more than one causal signal is present by comparing credible sets. Steps:

- Run **SuSiE** (Sum of Single Effects) fine-mapping on the region for each trait. The coloc package offers coloc.susie() which can integrate with SuSiE automatically. Make sure to provide an LD correlation matrix for SNPs in the region (calculated from a reference panel or the study if individual data available), using R:

```
# Ensure susieR is loaded
susie_out1 <- susie_rss(Bhat = exp_region$beta, Shat =
exp_region$se, R = LD_matrix, n = exp_sample_size, max_iter=100)
susie_out2 <- susie_rss(Bhat = out_region$beta, Shat =
out_region$se, R = LD_matrix, n = out_sample_size, max_iter=100)
susie_coloc <- coloc.susie(susie_out1, susie_out2, p1=1e-4, p2=1e-
4, p12=1e-5)
```

Here susie_rss performs fine-mapping given summary stats and LD, producing a set of credible sets for each trait. coloc.susie then compares these credible sets to compute a modified PP for colocalization, often denoted PP.H4* or similar.

- Interpret the SuSiE-based results: The output may show multiple "colocalizing" signal pairs if, say, each trait has 2 causal signals and one of them matches. For simplicity, you can take the maximum PP for any configuration as the evidence of colocalization. For example, if coloc.susie finds that one of the exposure signals colocalizes with one of outcome's, with PP=0.9, that is strong evidence. If using coloc.susie via the coloc package, it will report an integrated summary similar to coloc.abf. According to one approach, use the maximum PP.H4 from coloc.susie credible sets if multiple causal variants exist.
- Compare with the single-variant coloc from step 9. Ideally, both should be concordant. SuSiE can rescue cases where multiple causal variants made coloc.abf uncertain (e.g. high PP3). If SuSiE-coloc still shows no colocalization, it reinforces that exposure and outcome do not share signals at that locus.
- Fine-mapping also provides **credible sets of SNPs** for each trait's association. Overlap of these credible sets can pinpoint **candidate causal SNPs**. If a particular

SNP is in the 95% credible set of both exposure and outcome, it is likely the causal link. Note such SNP(s) as prioritized variants.

- Save the credible set info and SuSiE output if needed (perhaps in results/coloc/ with details on each signal).

4.11 **HAR annotation and prioritization:**

Now integrate the evolutionary annotation (HARs) and functional mapping:

- **Overlay HARs:** For each locus identified as having a potential causal relationship (especially those with colocalized signals), check if the instrument or lead SNP lies in or near a Human Accelerated Region. Using the HAR coordinate file (loaded as a GRanges or data.table of intervals), perform an overlap query. In R, one can do:

```
library(GenomicRanges)
har_regions <- import.bed("HAR_hg19.bed")  # or HAR_hg38.bed
depending on build
locus_gr <- GRanges(seqnames=chr, IRanges(start=lead_pos,
end=lead_pos))
overlap <- findOverlaps(locus_gr, har_regions)
```

Alternatively, check if any SNP in the locus credible set (if fine-mapped) falls in a HAR. Consider also a proximity threshold (e.g. within 50kb of a HAR) if none directly overlap, since HARs might be enhancer elements not exactly at the SNP.

- If an overlap is found, note the HAR ID. This suggests the association might involve a region of the genome that experienced rapid human-specific evolution.
- **Map to nearest genes:** Identify genes in the locus that could be mediators. This includes genes closest to the lead SNP, genes that contain the SNP (if exonic), or genes linked via eQTL. Use an annotation package or simply a gene list: for the lead SNP coordinate, query a gene annotation for any gene whose transcription start site is within e.g. ±100kb. Note these gene names.
- **Brain eQTL lookup:** For each candidate gene from the above step (or if specific SNP is known to be an eQTL for a gene), check if the variant is an expression quantitative trait locus in brain tissue. Query GTEx: if a variant is strongly associated ($p < 1e-5$) with expression of gene X in cortex, that strengthens a hypothesis that the variant influences the outcome via gene X in brain. If you have a pre-fetched list of brain eQTLs, filter it for the SNP or region. Alternatively, use an online lookup (GTEx portal or QTL database) for top candidate SNPs. Document any significant eQTL relationships.
- **HAR-gene mapping:** If a HAR overlaps, see if that HAR was previously linked to regulating a specific gene (some HARs in literature are known enhancers for particular genes). If available from literature, mention that. This can be found in Pollard's HAR papers or by seeing which gene is nearest the HAR.
- The HAR-aware lens helps narrow which findings might be more biologically compelling (given HARs often tag developmental regulatory elements). However, do

not exclude non-HAR hits, they are also important, the HAR is just an added layer of interest.

4.12 **Multiple testing correction and reporting:**

If you tested multiple exposures and outcomes, adjust for multiple comparisons to control false discovery rate. If 10 exposures × 5 outcomes = 50 tests were done for the primary MR, calculate an FDR-adjusted q-value for each MR p-value. Use p.adjust(pvals, method="fdr") in R. Determine which results are significant at q < 0.05. These would be the robust findings to focus on. Mark them in the results table. If only a few tests, Bonferroni might be fine too (but FDR is less conservative and more appropriate if tests are not independent).

Additionally, within each exposure if it had multiple instruments (multiple loci), and you did colocalization for each, you might consider the chance of false positives in colocalization. It's less standard to adjust those probabilities, but you might set a stringent threshold like PP4 > 0.9 for claiming colocalization if many loci are tested. Usually, controlling FDR at the level of MR tests is primary.

Compile a comprehensive report of results:

- **Primary table:** Each exposure-outcome pair with MR results (IVW estimate, CI, p, number of SNPs, and FDR q). Indicate which remain significant after FDR. Include Egger results (especially intercept p for pleiotropy) in a separate column or separate table for sensitivity.
- **Colocalization table:** For each significant exposure–outcome (or each locus), list the locus, lead SNP, PP.H4, conclusion (colocalized yes/no).
- **Annotated findings:** The prioritized loci with HAR/eQTL info as prepared in step 11. Ensure all results are clearly linked to the hypotheses.
- Apply interpretation limits: Emphasize that MR indicates *consistent with causality* but not proof, especially for microbiome, where instruments explain tiny variance and horizontal pleiotropy is a risk. If any results rely on suggestive instruments (p<1e-5), note they are exploratory. If colocalization didn't confirm an MR finding, mention that as a caution (it could be a false positive MR or pleiotropy).

4.13 **Extended/alternative analyses (optional):**

Depending on study goals, additional analyses can be appended:

- **Polygenic Risk Score variance check:** If many instruments, compute the proportion of variance in outcome explained by the MR (using method of Burgess et al.) or use *MVMR* (multivariable MR) if multiple exposures are correlated (e.g. microbiome features co-occurring). Not detailed here, but mention if relevant.
- **Stratified analysis by ancestry or sex:** If outcome data by subgroups exist, run MR in each to see consistency.
- **Phenome-wide MR (hypothesis free):** Test one exposure against a range of outcomes (or vice versa) to see if the exposure has effects on other traits (but then multiple testing correction becomes very stringent).

- **Pleiotropy assessments:** Use MR-PRESSO results to identify which SNPs were outliers; investigate those SNPs in GWAS Catalog to see if they have known associations that could confound (e.g. a microbiome SNP also affects immune diseases, then outcome effect might be via immunity not via microbiome).
- **Visualization:** Create a summary figure of the pipeline or results (see Appendix figure captions). A circos or network plot linking microbes to outcomes where causal links were found (microbe nodes to disease nodes).
- **Replication attempts:** If an independent dataset is available for either exposure or outcome, attempt replication MR. E.g., if an outcome has a second GWAS or a subset like males vs females. Or use a different microbiome GWAS (if one exists) to replicate the association. These are not required for core inference but add robustness. Each would follow similarly structured steps.

4.14 **Finalize outputs and documentation:**
Double-check that all results files are saved and backed up. Output the session information for reproducibility:

```
writeLines(capture.output(sessionInfo()), "logs/sessionInfo.txt")
```

This logs R version and all package versions used. Save your R script or notebook as the computational record.

Protocol references

- Kurilshikov, A., Medina-Gomez, C., Bacigalupe, R., Radjabzadeh, D., Wang, J., Demirkan, A., Le Roy, C. I., Raygoza Garay, J. A., Finnicum, C. T., Liu, X., Zhernakova, D. V., Bonder, M. J., Hansen, T. H., Frost, F., Rühlemann, M. C., Turpin, W., Moon, J.-Y., Kim, H.-N., Lüll, K., ... Zhernakova, A. (2021). Large-scale association analyses identify host factors influencing human gut microbiome composition. Nature Genetics, 53(2), 156–165. https://doi.org/10.1038/s41588-020-00763-1
- Cheng, L., Qi, C., Yang, H., Lu, M., Cai, Y., Fu, T., Ren, J., Jin, Q., & Zhang, X. (2021). gutMGene: a comprehensive database for target genes of gut microbes and microbial metabolites. Nucleic Acids Research, 50(D1), D795–D800. https://doi.org/10.1093/nar/gkab786
- Zuber, V., Grinberg, N. F., Gill, D., Manipur, I., Slob, E. A. W., Patel, A., Wallace, C., & Burgess, S. (2022). Combining evidence from Mendelian randomization and colocalization: Review and comparison of approaches. The American Journal of Human Genetics, 109(5), 767–782. https://doi.org/10.1016/j.ajhg.2022.04.001
- Wallace, C. (2020). Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. PLOS Genetics, 16(4), e1008720. https://doi.org/10.1371/journal.pgen.1008720
- Wallace, C. (2021). A more accurate method for colocalisation analysis allowing for multiple causal variants. PLOS Genetics, 17(9), e1009440. https://doi.org/10.1371/journal.pgen.1009440
- Pollard, K. S., Salama, S. R., Lambert, N., Lambot, M.-A., Coppens, S., Pedersen, J. S., Katzman, S., King, B., Onodera, C., Siepel, A., Kern, A. D., Dehay, C., Igel, H., Ares, M., Jr, Vanderhaeghen, P., & Haussler, D. (2006). An RNA gene expressed during cortical development evolved rapidly in humans. Nature, 443(7108), 167–172. https://doi.org/10.1038/nature05113
- Whalen, S., Inoue, F., Ryu, H., Fair, T., Markenscoff-Papadimitriou, E., Keough, K., Kircher, M., Martin, B., Alvarado, B., Elor, O., Laboy Cintron, D., Williams, A., Hassan Samee, Md. A., Thomas, S., Krencik, R., Ullian, E. M., Kriegstein, A., Rubenstein, J. L., Shendure, J., ... Pollard, K. S. (2023). Machine learning dissection of human accelerated regions in primate neurodevelopment. Neuron, 111(6), 857-873.e8. https://doi.org/10.1016/j.neuron.2022.12.026
- Skrivankova, V. W., Richmond, R. C., Woolf, B. A. R., Yarmolinsky, J., Davies, N. M., Swanson, S. A., VanderWeele, T. J., Higgins, J. P. T., Timpson, N. J., Dimou, N., Langenberg, C., Golub, R. M., Loder, E. W., Gallo, V., Tybjaerg-Hansen, A., Davey Smith, G., Egger, M., & Richards, J. B. (2021). Strengthening the Reporting of Observational Studies in Epidemiology Using Mendelian Randomization. JAMA, 326(16), 1614. https://doi.org/10.1001/jama.2021.18236

- Whalen, S., Inoue, F., Ryu, H., Fair, T., Markenscoff-Papadimitriou, E., Keough, K., Kircher, M., Martin, B., Alvarado, B., Elor, O., Laboy Cintron, D., Williams, A., Hassan Samee, Md. A., Thomas, S., Krencik, R., Ullian, E. M., Kriegstein, A., Rubenstein, J. L., Shendure, J., ... Pollard, K. S. (2023). Machine learning dissection of human accelerated regions in primate neurodevelopment. Neuron, 111(6), 857-873.e8. https://doi.org/10.1016/j.neuron.2022.12.026