Mind the gap: characterizing bias due to population mismatch in two-sample Mendelian randomization

Jack Li[1*], Jean Morrison[1†]
1: University of Michigan, Department of Biostatistics
* lijack@umich.edu; †jvmorr@umich.edu

July 2025

1 Abstract

Mendelian randomization (MR) is a statistical method for estimating causal effects using genetic variants as instrumental variables. In two sample MR (2SMR), different study samples are used to estimate genetic associations with the exposure and outcome. For valid inference, these studies must include individuals from the same population. Using studies from different populations may bias the 2SMR estimate due to differences in linkage disequilibrium or genetic effects on the exposure trait. We show that violation of the same-population assumption leads to bias in the causal estimate towards zero on average, and does not increase the rate of false positives. We verify this result in a broad survey of 2SMR estimates, comparing estimates made with matching and mismatching populations across 546 trait pairs measured in 2-7 ancestries. We find that most population-mismatched estimates are attenuated towards zero compared to their corresponding population-matched estimates, and that increasing genetic distance between study populations is associated with greater shrinkage. We observe bias even when mismatched populations have the same continental ancestry. However, we also find that, in some cases, using a larger exposure study with mismatching ancestry can improve power by dramatically increasing precision. These results show that even intra-continental population mismatch can bias 2SMR estimates, but also suggests there is potential to improve the power of 2SMR in understudied populations by properly leveraging larger, mismatching study populations.

2 Introduction

Mendelian randomization (MR) is a popular statistical method for estimating the causal effect of an exposure on an outcome using genetic variants as instrumental variables [1]. MR has become extremely common in recent years due to its potential to estimate causal effects using only summary information for genetic associations, which are readily available through public resources such as the OpenGWAS Project and the EMBL-EBI GWAS catalog [2, 3]. Two-sample Mendelian randomization (2SMR) is a variation of MR in which one sample is used to estimate variant-exposure associations and a second sample is used to estimate variant-outcome associations. These two components are then combined into a single MR estimate. 2SMR is the most popular form of MR because it enables estimation of a wider range of causal effects than one-sample MR (1SMR) by not requiring that both exposure and outcome are measured in the same sample. Additionally, using many common estimation methods, 2SMR is preferable to 1SMR even when the latter is possible, because 2SMR estimates are biased towards the null while 1SMR estimates are biased towards the confounded population association [4]. However, more advanced methods such as GRAPPLE[5], CAUSE[6], and MR-BEE[7] can account for this bias.

Valid estimation using 2SMR relies on the assumption that the exposure and outcome GWAS studies are sampled from the same underlying population, known as the same-population assumption. Haycock et al. [1] note that even though it may be possible to test the hypothesis of no causal effect when the same-population assumption is violated, the magnitude of the causal estimate may not be accurate. Moreover, Burgess et al. [8] note that a variant may be valid in one population but not the other, and that more generally, an association in one population may not be replicated in another. This assumption is also reflected in the Strengthening the Reporting of Observational Studies in Epidemiology using Mendelian Randomization (STROBE-MR) guidelines, which recommend that authors describe the study design and underlying population of each dataset [9].

Typically, investigators conducting 2SMR attempt to fulfill the same-population assumption by selecting GWAS with the same continental ancestry. For example, a multi-ancestry 2SMR study of the effect of the proteome on a common diseases used pQTL data from Europeans from ARIC and INTERVAL and outcome disease GWAS of Europeans from the Global Biobank Meta-Analysis Initiative (GBMI) [10]. However, populations used in 2SMR studies often do not match beyond continental-level ancestry. For example, researchers commonly use summary statistics from meta-analyses in 2SMR studies

due to their large sample sizes [11, 12, 13, 14, 15]. This practice guarantees some amount of population mismatch between exposure and outcome samples.

Although matching study populations on continental ancestry is common, it is unclear how effective this practice is at mitigating bias due to violation of the same-population assumption. Additionally, it is possible that this practice is too restrictive and inhibits performing MR studies of effects in populations for which there are few or under-powered population-matched exposure GWAS. It may also discourage researchers from making full use of large multi-ancestry meta analyses, potentially reducing power of the resulting analyses.

In this paper, we characterize how violations of the same-population assumption affect 2SMR estimates. First, we show theoretically that violations of the same-population assumption are expected to bias the 2SMR causal effect estimate towards zero using the most common instrument selection strategy. Second, we compare population matched and population mismatched 2SMR estimates across a wide range of traits and ancestries to systematically estimate bias due to population differences. In agreement with our theoretical expectations, we find that population mismatch nearly always attenuates signals of causal effects toward zero. Finally, we evaluate the trade-off between bias and power when choosing between a smaller population-matched exposure study and a large population-mismatched exposure study. Our results suggest that, in some cases, substantial power gain can be realized by violating the same-population assumption, at the cost of null-biased effect estimates.

3 Methods

3.1 Effect of Population Mismatch on 2SMR Estimates

Figure 1 illustrates the design of a 2SMR study, as well as the components of standard 2SMR effect estimates. A 2SMR study makes use of two different GWAS studies, one for the exposure, X, performed in population e (the exposure population) and one for the outcome, Y, performed in population t (the target population). Instruments for the exposure, $G_1, \ldots, G_J$, are typically selected if they are strongly associated with X in the exposure population. Finally, GWAS associations for each instrument with the exposure and outcome are used to estimate the causal effect.

Let γ_t be the average causal effect of X on Y in population t. All MR methods rely on three core assumptions, all assumed to hold in population t: (i) G_j is associated with X, (ii) there are no confounders of G_j and the outcome, and (iii) G_j is only associated with Y through its association with X. If these assumptions are satisfied, then

$$\beta_{Y,j,t} = \gamma_t \beta_{X,j,t}, \tag{1}$$

where $\beta_{Y,j,t}$ is the association of variant j with Y in the target population and $\beta_{X,j,t}$ is the association of variant j with X in the target population. Some methods extend equation (1) to $\beta_{Y,j,t} = \gamma_t \beta_{X,j,t} + \alpha_j$ where α_j captures horizontal pleiotropic associations of G_j with Y (violations of assumption (iii)), and is usually assumed to follow a known distributional form. If both GWAS are performed in samples from population t, we can obtain estimates, $\hat{\beta}_{Y,j,t}$ and $\hat{\beta}_{X,j,t}$ of $\beta_{Y,j,t}$ and $\beta_{X,j,t}$, respectively. Methods such as inverse-variance weighted (IVW) regression [16] and the weighted median estimator [17] compute estimates, $\hat{\gamma}_t$, of γ_t as a weighted average of the ratio $\frac{\hat{\beta}_{Y,j,t}}{\hat{\beta}_{X,j,t}}$, as shown in Figure 1. Other methods such as GRAPPLE [5] or CAUSE [6] estimate γ_t are likelihood-based, but still rely on equation (1) or its extension including the term for horizontal pleiotropy.

When GWAS estimates are obtained in two different populations, the same-population assumption is additionally required. This assumption ensures that $\beta_{X,j,e}$, the association between variant G_j and X in the exposure population, is equal to $\beta_{X,j,t}$ and, therefore $\beta_{Y,j,t} = \gamma_t \beta_{X,j,t} = \gamma_t \beta_{X,j,e}$. If the causal effect of X on Y differs between the exposure and target populations, but $\beta_{X,j,e} = \beta_{X,j,t}$ holds for all j, then 2SMR estimates γ_t, the causal effect in the target population and not γ_e, the causal effect in the exposure population.

The assumption that $\beta_{X,j,e} = \beta_{X,j,t}$ can fail in two ways, either through gene-environment interactions or differences in linkage disequilibrium (LD) patterns, as illustrated in Figure 1. The true marginal associations, $\beta_{X,j,e}$ and $\beta_{X,j,t}$, can be expressed as

$$\beta_{X,j,\star} = \sum_{i=1}^{n_j} r_{i,j,\star} \delta_{X,i,\star} \qquad \star \in \{t, e\} \tag{2}$$

where $\delta_{X,i,\star}$ is the causal effect of variant i on X in population $\star$, $r_{i,j,\star}$ is the correlation between variant i and variant j in population $\star$, and the sum is over all n_j variants in LD with variant j. Gene-environment (or gene-gene) interactions, result in differences in the true causal genetic effects, $\delta_{X,i,t}$ and $\delta_{X,i,e}$, while differences in LD correspond to differences between $r_{i,j,t}$ and $r_{i,j,e}$. If either or both of these components differ across populations then $\beta_{X,j,e} \neq \beta_{X,j,t}$.

Since variants are typically selected based on their strength of association with X in the exposure population, $\beta_{X,j,e}$ is, on average, larger in magnitude than $\beta_{X,j,t}$ among the selected variants. This form of winner's curse is well known and documented. Xiao and Boehnke [18] observe that GWAS associations are typically smaller in magnitude in a replication cohort than a discovery cohort and Palmer and Pe'er [19] note that this effect is stronger when discovery and replication

cohorts have different ancestries. As a result, in the presence of a nonzero causal effect, we expect the ratio estimator $\frac{\hat{\beta}_{Y,j,t}}{\hat{\beta}_{X,j,e}}$ to be closer to zero than $\frac{\hat{\beta}_{Y,j,t}}{\hat{\beta}_{X,j,t}}$. Therefore, in the presence of a true causal effect, 2SMR estimates obtained with mismatching populations are expected to be biased towards zero. However, this is not guaranteed to be the case for all trait and population combinations because shrinkage in variant-exposure associations is only on average and not uniform across instruments.

By contrast, in the absence of a true causal effect, $\beta_{Y,j,t}$ is equal to zero as long as the selected variant does not exhibit horizontal pleiotropy in the target population. This means that the expectation of the ratio estimator is 0 even if $\beta_{X,j,e}$ differs from $\beta_{X,j,t}$. Therefore, if there is no horizontal pleiotropy, using a population-mismatched exposure study results in an unbiased estimate of the null causal effect. Horizontal pleiotropy is the existence of an association between a selected variant and Y in the target population through mechanisms other than a causal effect of X on Y. Variants are selected based on the strength of their association with X only, so presence or absence of horizontal pleiotropy does not affect the chance that a variant is selected as an instrument, regardless of which population is used for selection. This means that violation of the same-population assumption does not increase the risk of false positives due to horizontal pleiotropy.

3.2 Empirical Survey of the Impact of Population Mismatch on 2SMR Estimates

3.2.1 Study Design

To empirically assess the effect of population mismatch on 2SMR estimates, we compared 2SMR estimates made in matching populations to those made in mismatching populations across a large number of traits and populations, as illustrated in Figure 2. We selected 26 exposure traits including anthropometric, clinical, and molecular measurements. For each exposure trait, we identified one outcome trait with prior evidence of a causal link, resulting in a total of 21 outcome traits, as some exposures were linked to the same outcome. These include 20 binary diseases and one continuous trait, total billirubin. This strategy ensures that the set of all pairwise combinations of exposure and outcome traits includes some expected non-null effects as well as some likely null effects. Exposure and outcome traits are listed in Supplementary Tables S1 and S2.

We identified publicly available GWAS summary statistics for all exposure and outcome studies from as many populations as possible using the OpenGWAS database, the EMBL-EBI GWAS catalog, and FinnGen Release 10 [2, 3, 20]. We identified an average of 5 studies per exposure trait (128 total exposure studies) and 4 studies per outcome trait (88 total outcome studies). In total, there are 546 exposure-outcome trait pairs, and 11,392 exposure-outcome study pairs (approximately 21 study pairs per trait pair). The study source for each trait is given in Supplemental Data 1.

2SMR estimates the causal effect in the target population, which is the population in which the outcome GWAS is performed. Therefore, we compared 2SMR estimates within $88 \times 26 = 2,288$ groups defined by an exposure-outcome pair and a target population (Figure 2). This ensures that any differences observed in MR estimates are due to differences in the exposure population and not population differences in the causal effect.

Different studies of the same exposure may report effects in different units. In order to obtain comparable 2SMR effect size estimates from different exposure studies, we converted all exposure studies for the same trait to a standard unit, such as mmol/L for LDL cholesterol. In the case of studies that were inverse-normalized or z-score transformed, we scaled variant effect estimates by the standard deviation of the exposure trait if it was listed, or used an external estimate when feasible. Studies in which the exposure trait was nonlinearly transformed (e.g., log or square root) and meta-analyses that combine estimates on different scales could not be converted to a standard unit but were retained for comparisons of sign and significance of causal estimates. In total, we were able to convert 89 of the 128 exposure studies to a standard unit (Supplemental Data 2). Estimates for binary outcome were all either reported on the log odds scale or converted from the standardized scale to the log odds scale by multiplying by $\mu(1-\mu)$, where μ is the proportion of cases in the study cohort (Supplemental Data 3).

3.2.2 Two-Sample Mendelian Randomization

We performed 2SMR across all exposure-outcome study pairs using IVW regression and GRAPPLE [5]. For both estimators, instruments were selected from the exposure GWAS after LD clumping, prioritizing variants with the strongest exposure association with threshold $r^2 < 0.01$. This practice can result in a small amount of bias due to winner's curse [21] but is the most common approach used in 2SMR studies. For IVW regression, instruments were selected using a p-value threshold of 5×10^{-8}. For GRAPPLE, we used two thresholds, 5×10^{-8} and 1×10^{-5}, as suggested by the authors of the GRAPPLE method, which is more robust to weak instrument bias. We additionally performed Steiger filtering to remove variants more strongly associated with the outcome than the exposure.

We computed each estimator for all exposure-outcome study pairs with at least two instruments. Using 5×10^{-8} as the instrument selection p-value threshold, we could analyze $9,828$ pairs, of which 7,239 are on a standard scale. Using the 1×10^{-5} threshold, we could analyze $11,248$ pairs, of which 7,829 are on a standard scale. Complete details of our implementation of Steiger filtering and both MR estimation methods are given in Supplementary Note Section 8.1.

We compared 2SMR estimates made in matching and mismatching populations, under the assumption that differences in these estimates are due to violation of the same-population assumption. However, this assumption is not always justifiable using the IVW regression method, due to sample overlap in the population-matched analysis. When MR is performed in overlapping samples, weak-instrument bias in IVW regression is towards the observational association, rather than towards

zero [4]. GRAPPLE estimates do not suffer from this problem because sample overlap is accounted for using a pre-computed estimate of the residual correlation between variant-exposure and variant-outcome association estimates [5]. Therefore, we present results from GRAPPLE as our primary analysis and include results using IVW regression in the Supplemental Material.

3.2.3 Quantifying the effect of population mismatch on 2SMR

Of the 2,314 groups defined by an exposure-outcome trait pair and a target population, exactly matching exposure studies that can be converted to a standardized scale are available for 818 groups at an instrument selection p-value threshold of 1×10^{-5}, and 813 at an instrument selection p-value threshold of 5×10^{-8} (Figure 2). We classify an exposure study is an exact match if it is sampled from the same study cohort as the outcome. For each of these groups, we designate the 2SMR estimate using the matching exposure population as the reference estimate, $\hat{\gamma}_r$.

We used two strategies to quantify bias due to violation of the same-population assumption. First, for each population-mismatched exposure study, we computed a z-score to test the hypothesis that the population-mismatched and reference estimate have the same expectation,

$$Z_{concord} = \frac{(\hat{\gamma}_{mm} - \hat{\gamma}_r)}{\sqrt{\sigma^2_{mm} + \sigma^2_r}}, \tag{3}$$

where $\hat{\gamma}_{mm}$ is the population-mismatched estimate, and σ^2_{mm} and σ^2_r are estimated standard errors of $\hat{\gamma}_{mm}$ and $\hat{\gamma}_r$. If $\hat{\gamma}_{mm}$ and $\hat{\gamma}_r$ have the same expectation, then $Z_{concord}$ follows a standard normal distribution. For each z-score we computed a p-value testing the hypothesis that the z-score was drawn from the null distribution. We used the Benjamini-Hochberg procedure to compute estimates of the false discovery rate (FDR) within estimate groups for each instrument selection threshold. We classified mismatching estimates that were significantly different from their reference estimate at FDR < 0.05 into three categories: estimates with the same sign as the reference estimate and closer to zero, estimates with the same sign as the reference estimate and farther from zero, and estimates with the opposite sign as the reference estimate.

Second, to quantify the average degree of shrinkage or inflation resulting from population mismatch, we regressed mismatching estimates on corresponding reference estimates with no intercept, weighted by the inverse-variance of the mismatching estimates. We used simulation extrapolation (SIMEX) to account for uncertainty in the 2SMR estimates (Supplementary Note Section 8.2) [22]. The resulting estimated slope provides an estimate of the expected shrinkage or inflation due to violation of the same-population assumption. An estimate less than 1 indicates that estimates obtained using mismatching populations tend to be closer to zero than estimates obtained using matching populations. We used bootstrapping with 1,000 replicates to estimate the standard error of these SIMEX-estimated shrinkage coefficients.

We computed the SIMEX regression slope across all analyses with available reference estimates and within groups defined by exposure-outcome population pairs. At an instrument selection p-value threshold of 1×10^{-5}, we identified 26 population pairs with at least eight estimates with an available reference estimate. This threshold was selected to ensure that we could calculate SIMEX regression slopes for at least some population pairs using African populations as a target, while maximizing the precision of any estimates we calculate. At the stricter selection p-value threshold of 5×10^{-8}, 23 population pairs have at least eight estimates with an available reference estimate. However, none of these pairs include African populations as a target population. We expect greater shrinkage in population pairs with larger differences in LD. To assess this prediction, we regressed the SIMEX estimated slope against the approximate F_{st} distance between the populations, obtained from Scott et al. [23] (Supplementary Table S3), weighted by the inverse of the variance of each estimate. The SIMEX regression procedure does not account for correlation between estimates that share the same exposure or outcome studies. This correlation may lead to under-estimation of standard errors, motivating our use of the bootstrap to estimate standard errors of shrinkage coefficients.

3.3 Trade-off between bias and power

To assess the trade-off between bias and power when using a larger mismatching exposure study, we examined 195 causal estimates made using outcomes measured in Biobank Japan. We compared population-matched estimates with estimates obtained using exposures from a meta analysis of UK Biobank and Biobank Japan [24], which is nearly three times as large as Biobank Japan alone. As the meta-analysis study combined effects of two populations measured on different scales, the population-mismatched MR estimates computed using the meta-analysis as exposure could not be converted to the standard scale. Therefore, we focus primarily on the sign and significance of these effect estimates.

We identified trait pairs that were significant in the population-mismatched analysis but non-significant in the population-matched analysis and vice-versa using a Bonferroni corrected signficance threshold of $p < 0.05/195$. We assessed the replicability of effects found in only one of the two analyses by comparing results with estimates from population-matched analyses using samples that do not overlap with either UK Biobank or Biobank Japan. The significance threshold for replication was set at the Bonferroni adjusted 0.05 threshold accounting for the total number of effects we attempted to replicate.

4 Results

4.1 Using a mismatching exposure biases 2SMR estimates towards zero

Figure 3 shows GRAPPLE MR estimates for the effect of LDL cholesterol on coronary heart disease in four target populations. Reference estimates are available for Biobank Japan and UK Biobank target populations. There are no reference estimates when using FinnGen or the CARDioGRAMplusC4D studies to measure the outcome, but there are estimates made using exposure studies with majority matching continental ancestry. MR estimates made using mismatching populations are all attenuated towards zero compared to the reference or continentally matched estimates. Using the instrument selection p-value threshold of 1×10^{-5}, all population-mismatched estimates are significantly closer to zero than the reference estimates at an FDR threshold of 5% based on the concordance z-score. We observe the same pattern at the stricter selection threshold, though differences are less significant due to higher variance in the causal effect estimates.

The shrinkage shown in Figure 3 is exemplary of a general pattern across all trait pairs. The SIMEX-estimated overall shrinkage coefficient is significantly less than 1 at both instrument selection thresholds (0.41 (SE 0.015) for 1×10^{-5} threshold and 0.46 (SE 0.019) for 5×10^{-8} threshold). At both instrument selection thresholds, the majority of population-mismatched estimates with significant concordance z-scores were of the same sign and closer to zero than their corresponding reference estimates (285/404 (71%) for 10^{-5} threshold, 105/160 (66%) for 5×10^{-8} threshold). In the majority of the remaining cases, the population-mismatched estimate has opposite sign from the reference estimate. However, in most of these cases (113/118 (96%) for 10^{-5} threshold, 43/50 (86%) for 5×10^{-8} threshold) either the population-mismatched or reference estimate is not significantly different from zero at a nominal 0.05 significance threshold. We found only one (10^{-5} threshold) and five (5×10^{-8} threshold) instances of a significant concordance z-score and a population-mismatched estimate that is of the same size and larger in magnitude than the corresponding reference estimate (Supplementary Table S5). We found similar results using IVW regression (Supplementary Table S4). However, we note that IVW regression reference estimates may be biased due to sample overlap.

Full details of all 7,829 and 7,239 standardized 2SMR estimates with and without Steiger filtering at both instrument selection thresholds using GRAPPLE and at the stringent selection threshold using IVW are provided in Supplemental Data 5, 6, and 7. The use of Steiger filtering did not substantially impact most MR estimates or our main findings (Supplementary Figure S1).

4.2 Average shrinkage due to population-mismatch is correlated with Fst

Figure 4 shows population pair-specific SIMEX regression estimates (shrinkage coefficients) vs approximate F_{st} distance between populations using the selection threshold 1×10^{-5} results. We observe a clear negative relationship between regression estimates and F_{st}, supported by inverse-variance weighted regression of shrinkage coefficients against F_{st}, which gives a regression coefficient of -2.40 (SE 0.56, p-value 0.0004). However, we also note that F_{st} does not completely determine the amount of shrinkage for all population pairs and there is a lack of symmetry in shrinkage coefficients between population pairs. For example, the shrinkage coefficient for UK Biobank used as exposure and FinnGen defining the target population is 0.67 (SE 0.03), while measuring the exposure in FinnGen and outcome in UK Biobank results in a shrinkage coefficient of 0.52 (SE 0.02). Notably, the shrinkage coefficient is significantly less than one for all population pairs, even those with matching continental ancestry (Supplemental Figure S2). Results are similar using the 5×10^{-8} variant selection threshold, though these do not include pairs involving African populations (Supplemental Figure S2 and Supplementary Note Section 8.3).

4.3 Large mismatching exposure studies may increase power of 2SMR

Of the 195 causal estimates computed using Biobank Japan as both exposure and outcome population, 13 are significantly different from zero at $p < 0.05/195$ using the more stringent instrument selection threshold, and 17 are significantly different from zero using the more lenient threshold. By contrast, when using a meta analysis of UK Biobank and Biobank Japan as the exposure study, 25 and 26 estimates are significantly different from zero at the more and less stringent instrument selection thresholds, respectively. Using the larger exposure study results in 14 significant causal effects that are not discovered using the population-matched exposure study at both instrument selection thresholds, while 2 and 5 causal estimates are significant in the population-matched analysis only at the more and less stringent instrument selection thresholds, respectively (Table 2 and Supplemental Data 8).

We checked if each causal estimate that differed in significance in the population-matched and population-mismatched analyses replicated in population-matched estimates for non-UK and non-Japanese populations. We were able to perform replication analysis on 13 of the 19 causal effects that differed in significance at the 1×10^{-5} threshold, and 12 of the 16 trait pairs that differed in significance at the 5×10^{-8} threshold. At both instrument selection thresholds, all estimates that were significant only using BBJ alone as the exposure study replicated in at least one other population. At the more lenient instrument selection threshold, 8 out of 10 trait pairs identified as significant only using the meta-analysis as the exposure study exposure showed evidence of replication (see Supplementary Table S6). At the more stringent selection threshold, only 5 out of the 11 trait pairs identified as significant only using the meta analysis showed evidence of replication. However, five

out of six of these non-replicating causal effects at the more stringent threshold had only estimates from studies with fewer than 50,000 participants available for replication analysis (Supplementary Table S7 and Supplemental Data 9).

5 Discussion

We demonstrate, both theoretically and through a comprehensive empirical survey, that bias in 2SMR estimates due to population mismatch usually results in a reduction in magnitude of the estimated effect. This bias occurs even when using continentally matching populations, a common practice in existing 2SMR literature. Over approximately 2,000 comparisons, we found fewer than a dozen instances where population-mismatched estimates were significantly different from their reference estimate and either larger in magnitude or opposite sign with both estimates nominally significantly different from zero. This suggests that anti-conservative bias due to population mismatch is extremely rare.

Additionally, we demonstrate a bias-variance tradeoff between population mismatch and sample using the example of outcomes measured in Biobank Japan and exposure studies sourced either from Biobank Japan or from a meta analysis of UK Biobank and Biobank Japan. We show that using a larger but population-mismatched exposure study may improve power compared to a smaller, population-matched study, at the expense of some bias towards zero in causal estimates. We argue theoretically that estimates made with population-mismatched studies are not more likely to be false positives than those made using matching studies, though both may be subject to bias due to horizontal pleiotropy if this is not appropriately accounted for. This result suggests that there may be utility in violating the same-population assumption in 2SMR analysis of effects in understudied populations. However, researchers should use caution when interpreting the magnitudes of these estimates. This result also motivates the development of more advanced statistical techniques that can be used to integrate and debias cross-population 2SMR estimates.

We observed bias at both more and less stringent instrument selection thresholds. However, bias was less pronounced at the more stringent threshold. We hypothesize that a more stringent threshold increases the proportion of instruments that are causal variants, partially resolving discrepancies due to differences in LD. This observation is supported by recent literature proposing the use of cross-ancestry fine-mapping [25, 26] to identify causal variants. However, we expect that even perfect causal variant identification will not eliminate bias due to population-mismatch because it does not address effect differences due to gene-environment interactions, and LD differences may still play a role when multiple causal variants are in LD in at least one population.

One limitation of our study results from the disparity in the abundance and quality of GWAS across populations. While we were always able to find at least one European ancestry study for each exposure and outcome trait, this was not true for other populations. For example, there is no available study of Vitamin D, multiple sclerosis, or Alzheimer's disease in an East Asian population in either OpenGWAS or the GWAS catalog. Summary statistics were even more limited in non-European, non-East Asian populations, including South Asian and African populations. Additionally European ancestry studies have sample sizes between 300,000 and 600,000, while the best or only available studies for South Asian and African populations were typically less than 10,000. As a result, our estimates of population mismatch-induced shrinkage in effect estimates is much less precise for these populations. Data availability for non-European ancestries also limited the power of our replication analysis in Section 4.3. However, even with improved parity in study availability and sample size between populations, we do not expect this limitation to affect the qualitative conclusion that mismatch typically leads to bias towards the null in effect estimates.

We found a significant, negative association between population-pair specific shrinkage coefficient estimates and F_{st} distance. However, there is heterogeneity in the degree of shrinkage across trait pairs. We also found that shrinkage due to population mismatch is not necessarily symmetric when exposure and outcome populations are swapped. Differences in the amount of shrinkage observed across trait pairs may be due both to differences in LD discrepancies around instruments and to differences in gene-environment interactions across traits. In this analysis, we have not characterized the relative contributions of these two factors.

6 Acknowledgements

This research was supported by National Human Genome Research Institute (NHGRI) R01 HG013104 and T32-HG000040. We want to acknowledge the participants and investigators of the FinnGen study for access to FinnGen summary data. We thank the work of the GWAS Catalog team and MRC Integrative Epidemiology Unit for making full summary statistics for GWAS studies publicly available through the GWAS Catalog and OpenGWAS projects, respectively. We thank Dr. Michael Boehnke, Dr. Laura Scott, Dr. Sebastian Zoellner, Dr. Hyun Min Kang and other members of the Center of Statistical Genetics at the University of Michigan for their helpful discussions and comments on the manuscript.

7 Figures

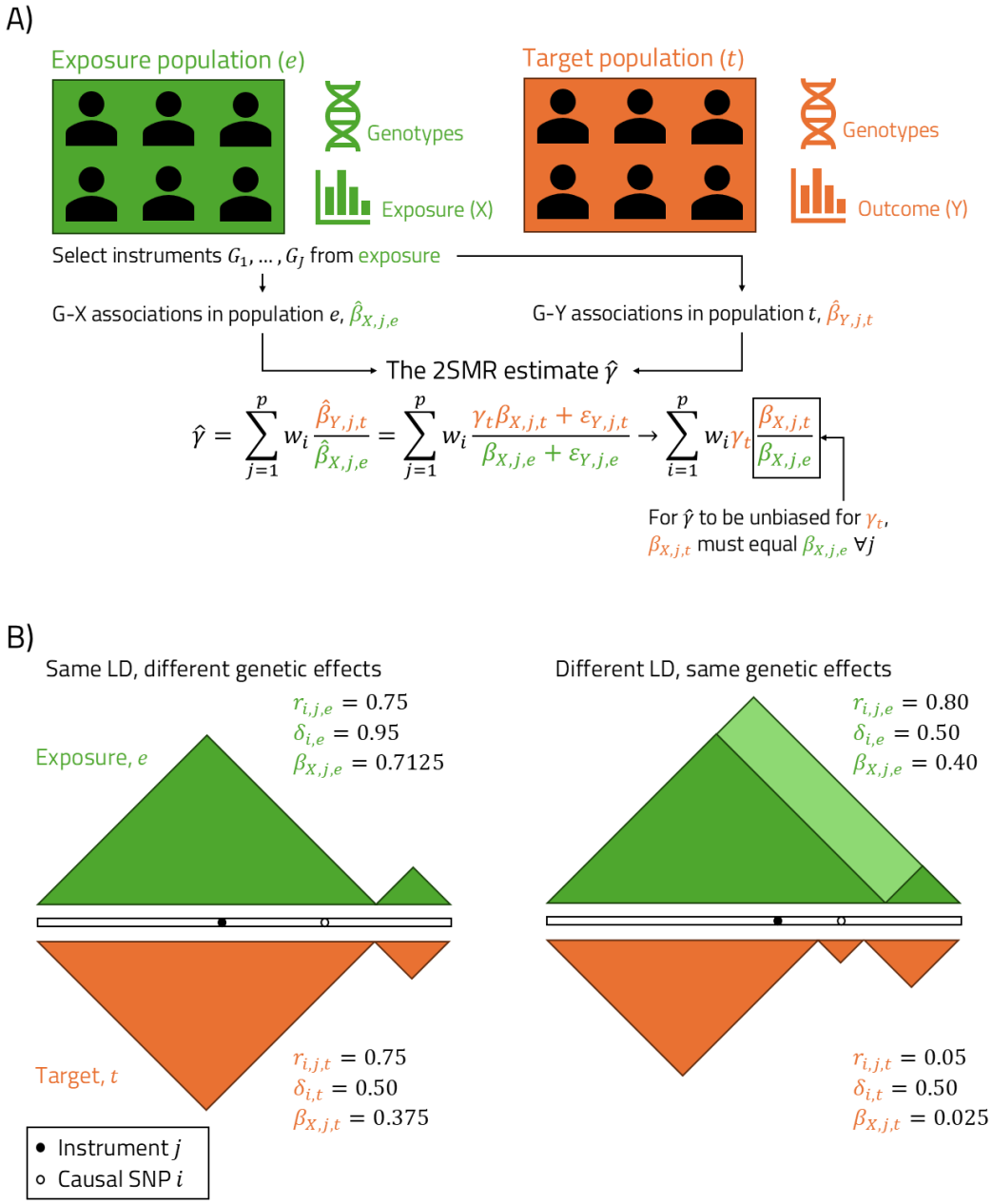


Figure 1: (A) Diagram depicting a 2SMR study and the calculation of the causal estimate as a weighted average of ratio estimates. (B) Two possible sources of differences between $\beta_{X,j,t}$ and $\beta_{X,j,e}$. Left, LD patterns are the same in the two populations but the causal variant has a different effect on the exposure. Right, the causal variant has the same effect in both populations, but LD structure differs. In both cases, there is a strong exposure-instrument association in population e, but a weaker association in population t.

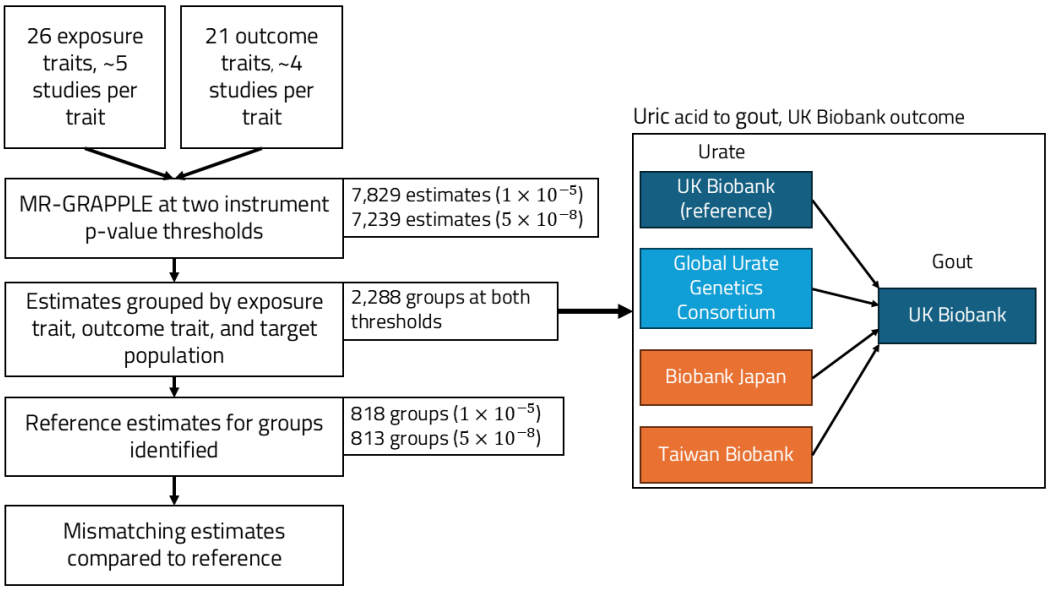


Figure 2: Study design of cross-population survey of 2SMR estimates. Right, example of an exposure trait-outcome study group for the effect of urate on gout with UK Biobank defining the target population. Exposure studies are color-coded based on how closely study populations match the target population: exact match (dark blue), continental match but a different subpopulation (light blue), or a continental mismatch (orange).

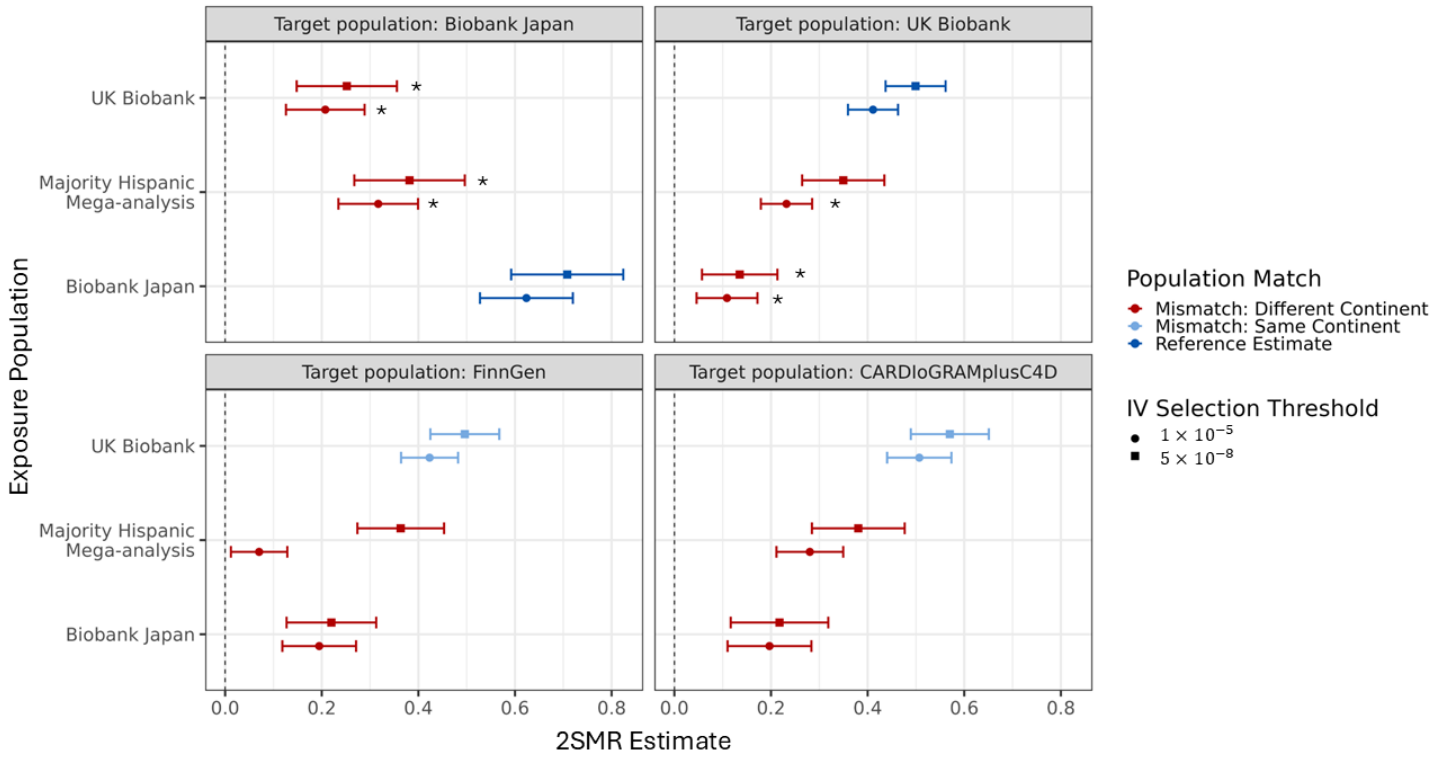


Figure 3: Forest plots of 2SMR estimates of the effect of LDL cholesterol on coronary heart disease from GRAPPLE at two instrument selection p-value thresholds. Points indicate effect estimates and intervals indicate 95% confidence intervals. Color indicates if an estimate is made with exact matching (dark blue), continentally matching (light blue), or continentally mismatching (red) exposure studies. A star indicates that an estimate is significantly different than the reference estimate (FDR < 0.05), when a reference estimate is available.

Instrument Selection Threshold	1×10^{-5}	5×10^{-8}
Total Number of Mismatching Estimates with Reference Estimates	2,182	1,951
Mismatching Estimates with Non-Significant Concordance z-scores	1,778	1,791
Mismatching Estimates with Significant Concordance z-scores, Same Sign as Reference, Smaller Magnitude	285	105
Mismatching Estimates with Significant Concordance z-scores, Same Sign as Reference, Larger Magnitude	1	5
Mismatching Estimates with Significant Concordance z-scores, Opposite Sign as Reference	118	50
Mismatching Estimates with Significant Concordance z-scores, Opposite Sign as Reference, Reference and Mismatch 2SMR p-values < 0.05	5	7

Table 1: A summary of how 2SMR estimates with mismatching exposures compared with their corresponding reference 2SMR estimates at two different instrument selection thresholds. All estimates were computed using GRAPPLE. Note that a lenient p-value threshold for instrument selection allowed for identification of instruments for more exposure studies, resulting in more mismatching estimates being available for analysis.

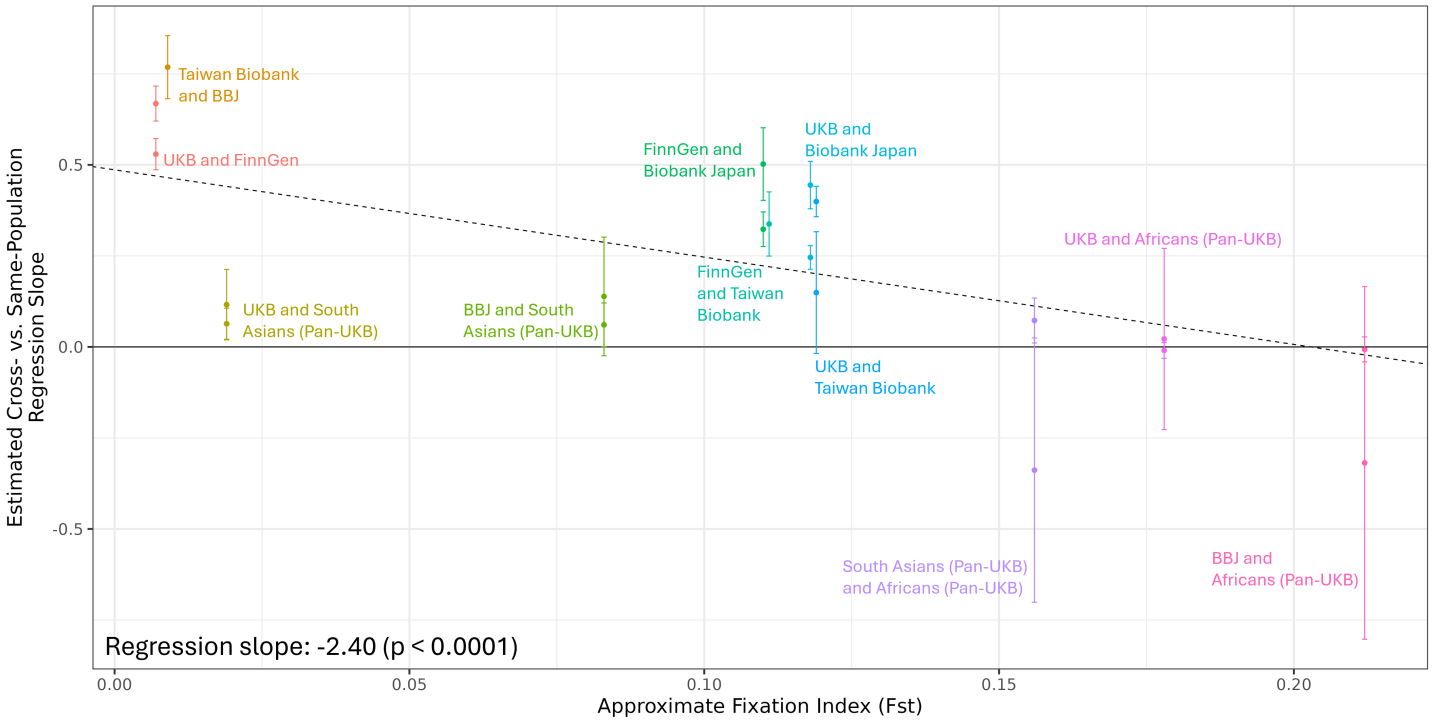


Figure 4: Population pair specific shrinkage coefficients vs approximate Fst distance between populations. Population pair specific shrinkage coefficients were estimated using SIMEX. The dashed line represents the univariable regression line from inverse-variance weighted regression of F_{st} on estimated shrinkage coefficient. Vertical confidence intervals represent 95% confidence intervals for each shrinkage coefficient.

	1×10^{-5} Threshold	5×10^{-8} Threshold
Total trait pairs assessed	195	195
Neither estimate significant	164	168
Both estimates significant	12	11
Only Meta-analysis to BBJ estimate significant	14	14
Replication Analyses Performed (Successful Replications), Only Meta-analysis to BBJ significant	10 (8)	11 (5)
Only BBJ to BBJ estimate significant	5	2
Replication Analyses Performed (Successful Replications), Only BBJ to BBJ significant	3 (3)	1 (1)

Table 2: Counts of significant and non-significant causal estimates with outcome studies from Biobank Japan (BBJ) and exposure studies from either BBJ or a meta analysis of BBJ and UK Biobank, as well as replication analysis results for trait pairs that were significant for only one choice of exposure. Significance of 2SMR estimates is based on a Bonferroni-corrected threshold of 0.05/195.

References

[1] Philip C Haycock, Stephen Burgess, Kaitlin H Wade, Jack Bowden, Caroline Relton, and George Davey Smith. Best (but oft-forgotten) practices: the design, analysis, and interpretation of mendelian randomization studies. *American Journal of Clinical Nutrition*, 103(4):965–78, 2016.

[2] Maria Cerezo, Elliot Sollis, Yue Ji, Elizabeth Lewis, Ala Abid, Karatuğ Ozan Bircan, Peggy Hall, James Hayhurst, Sajo John, Abayomi Mosaku, Santhi Ramachandran, Amy Foreman, Arwa Ibrahim, James McLaughlin, Zoë Pendlington, Ray Stefancsik, Samuel A Lambert, Aoife McMahon, Joannella Morales, Thomas Keane, Michael Inouye, Helen Parkinson, and Laura W Harris. The nhgri-ebi gwas catalog: standards for reusability, sustainability and diversity. *Nucleic Acids Research*, 53(D1):D998–D1005, 2025.

[3] Matthew S. Lyon, Shea J. Andrews, Ben Elsworth, Tom R. Gaunt, Gibran Hemani, and Edoardo Marcora. The variant call format provides efficient and robust storage of gwas summary statistics. *Genome Biology*, 22(32), 2021.

[4] Stephen Burgess, Neil M. Davies, and Simon G. Thompson. Bias due to participant overlap in two-sample mendelian randomization. *Genetic Epidemiology*, 40(7):597–608, 2016.

[5] Jingshu Wang, Qingyuan Zhao, Jack Bowden, Gibran Hemani, George Davey Smith, Dylan S. Small, and Nancy R. Zhang. Causal inference for heritable phenotypic risk factors using heterogeneous genetic instruments. *PLoS Genetics*, 17 (6), 2021.

[6] Jean Morrison, Nicholas Knoblauch, Joseph Marcus, Matthew Stephens, and Xin He. Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nature Genetics*, 52(7):740–747, 2020.

[7] Noah Lorincz-Comi, Yihe Yang, Gen Li, and Xiaofeng Zhu. Mrbee: A bias-corrected multivariable mendelian randomization method. *HGG Advances*, 5(3), 2024.

[8] Stephen Burgess, Robert A. Scott, Nicholas J. Timpson, George Davey Smith, Simon G. Thompson, and EPIC-InterAct Consortium. Using published data in mendelian randomization: a blueprint for efficient identification of causal risk factors. *European Journal of Epidemiology*, 30(7):543–552, 2015.

[9] Veronika W Skrivankova, Rebecca C Richmond, Benjamin A R Woolf, James Yarmolinsky, Neil M Davies, Sonja A Swanson, Tyler J VanderWeele, Julian P T Higgins, Nicholas J Timpson, Niki Dimou, Claudia Langenberg, Robert M Golub, Elizabeth W Loder, Valentina Gallo, Anne Tybjaerg-Hansen, George Davey Smith, Matthias Egger, and J Brent Richards. Strengthening the reporting of observational studies in epidemiology using mendelian randomization (strobe-mr) statement. *JAMA*, 326(16):1614–1621, 2021.

[10] Huiling Zhao, Humaria Rasheed, Therese Haugdahl Nøst, Yoonsu Cho, Yi Liu, Laxmi Bhatta, Arjun Bhattacharya, Global Biobank Meta analysis Initiative, Gibran Hemani, George Davey Smith, Ben Michael Brumpton, Wei Zhou, Benjamin M Neale, Tom R Gaunt, and Jie Zheng. Proteome-wide mendelian randomization in global biobank meta-analysis reveals multi-ancestry drug targets for common diseases. *Cell Genomics*, 2(11), 2022.

[11] Elsa Valdes-Marquez, Sarah Parish, Robert Clarke, Traiani Stari, Bradford B Worrall, METASTROKE Consortium of the ISGC, and Jemma C Hopewell. Relative effects of ldl-c on ischemic stroke and coronary disease. *Neurology*, 92(11): 1176–1187, 2019.

[12] Pengsheng Li, Haiyan Wang, Lan Guo, Xiaoyan Gou, Gengdong Chen, Dongxin Lin, Dazhi Fan, Xiaoling Guo, and Zhengping Liu. Association between gut microbiota and preeclampsia-eclampsia: a two-sample mendelian randomization study. *BMC Medicine*, 20(1):443, 2022.

[13] Yiwen Long, Lanhua Tang, Yangying Zhou, Shushan Zhao, and Hong Zhu. Causal relationship between gut microbiota and cancers: a two-sample mendelian randomisation study. *BMC Medicine*, 21(1):66, 2023.

[14] Fernando Pires Hartwig, Maria Carolina Borges, Bernardo Lessa Horta, Jack Bowden, and George Davey Smith. Inflammatory biomarkers and risk of schizophrenia: A 2-sample mendelian randomization study. *JAMA Psychiatry*, 74 (12):1226–1233, 2017.

[15] Holmes MV, Asselbergs FW, Palmer TM, Drenos F, Lanktree MB, Nelson CP, Dale CE, Padmanabhan S, Finan C, Swerdlow DI, Tragante V, van Iperen EP, Sivapalaratnam S, Shah S, Elbers CC, Shah T, Engmann J, Giambartolomei C, White J, Zabaneh D, Sofat R, McLachlan S; UCLEB consortium; Doevendans PA, Balmforth AJ, Hall AS, North KE, Almoguera B, Hoogeveen RC, Cushman M, Fornage M, Patel SR, Redline S, Siscovick DS, Tsai MY, Karczewski KJ, Hofker MH, Verschuren WM, Bots ML, van der Schouw YT, Melander O, Dominiczak AF, Morris R, Ben-Shlomo Y, Price J, Kumari M, Baumert J, Peters A, Thorand B, Koenig W, Gaunt TR, Humphries SE, Clarke R, Watkins H,

Farrall M, Wilson JG, Rich SS, de Bakker PI, Lange LA, Davey Smith G, Reiner AP, Talmud PJ, Kivimäki M, Lawlor DA, Dudbridge F, Samani NJ, Keating BJ, Hingorani AD, and Casas JP. Mendelian randomization of blood lipids for coronary heart disease. *European Heart Journal*, 36(9):539–550, 2015.

[16] Stephen Burgess, Adam Butterworth, and Simon G Thompson. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology*, 37(7):658–665, 2013.

[17] Jack Bowden, George Davey Smith, Philip Haycock, and Stephen Burgess. Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology*, 40(4):304–314, 2016.

[18] Rui Xiao and Michael Boehnke. Quantifying and correcting for the winner's curse in genetic association studies. *Genetic Epidemiology*, 33(5):453–462, 2009.

[19] Cameron Palmer and Itsik Pe'er. Statistical correction of the winner's curse explains replication variability in quantitative trait genome-wide association studies. *PLoS Genetics*, 13(7):e1006916, 2017.

[20] Mitja I. Kurki, Juha Karjalainen, Priit Palta, Timo P. Sipilä, Kati Kristiansson, Kati M. Donner, Mary P. Reeve, Hannele Laivuori, Mervi Aavikko, Mari A. Kaunisto, Anu Loukola, Elisa Lahtela, Hannele Mattsson, Päivi Laiho, Pietro Della Briotta Parolo, Arto A. Lehisto, Masahiro Kanai, Nina Mars, Joel Rämö, Tuomo Kiiskinen, Henrike O. Heyne, Kumar Veerapen, Sina Rüeger, Susanna Lemmelä, FinnGen, . . . , and Aarno Palotie. Finngen provides genetic insights from a well-phenotyped isolated population. *Nature*, 613:508–518, 2023.

[21] Eleanor Sanderson, M. Mario Glymour, Michael Holmes, Hyunseung Kang, Jean Morrison, Marcus Munafo, Tom Palmer, C. Mary Schooling, Chris Wallace, Qingyuan Zhao, and George Davey Smith. Mendelian randomization. *Nature Review Methods Primers*, 2(6), 2022.

[22] J.R. Cook and L.A. Stefanski. Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89(428):1314–1328, 1994.

[23] Eric M Scott, Anason Halees, Yuval Itan, Emily G Spencer, Yupeng He, Mostafa Abdellateef Azab, Stacey B Gabriel, Aziz Belkadi, Bertrand Boisson, Laurent Abel, Andrew G Clark, Greater Middle East Variome Consortium, Fowzan S Alkuraya, Jean-Laurent Casanova, and Joseph G Gleeson. Characterization of greater middle eastern genetic variation for enhanced disease gene discovery. *Nature Genetics*, 48(9):1071–1076, 2016.

[24] Saori Sakaue, Masahiro Kanai, Yosuke Tanigawa, Juha Karjalainen, Mitja Kurki, Seizo Koshiba, Akira Narita, Takahiro Konuma, Kenichi Yamamoto, Masato Akiyama, Kazuyoshi Ishigaki, Akari Suzuki, Ken Suzuki, Wataru Obara, Ken Yamaji, Kazuhisa Takahashi, Satoshi Asai, Yasuo Takahashi, Takao Suzuki, Nobuaki Shinozaki, Hiroki Yamaguchi, Shiro Minami, Shigeo Murayama, Kozo Yoshimori, Satoshi Nagayama, Daisuke Obata, Masahiko Higashiyama, Akihide Masumoto, Yukihiro Koretsune, FinnGen, Kaoru Ito, Chikashi Terao, Toshimasa Yamauchi, Issei Komuro, Takashi Kadowaki, Gen Tamiya, Masayuki Yamamoto, Yusuke Nakamura, Michiaki Kubo, Yoshinori Murakami, Kazuhiko Yamamoto, Yoichiro Kamatani, Aarno Palotie, Manuel A Rivas, Mark J Daly, Koichi Matsuda, and Yukinori Okada. A cross-population atlas of genetic associations for 220 human phenotypes. *Nature Genetics*, 53(10):1415–1424, 2021.

[25] Gibran Hemani, Yoonsu Cho, Amanda Chong, Tom Palmer, Amy Mason, John Ferguson, David Evans, and George Davey Smith. Jointly modelling multiple ancestral populations using gwas summary data improves causal inference. *biorXiv [PREPRINT]*, 2025.

[26] Yihe Yang and Xiaofeng Zhu. Improving causal effect estimation in multi-ancestry multivariable mendelian randomization with transfer learning. *bioRxiv*, 2025. doi: 10.1101/2025.07.11.664423.

[27] Gibran Hemani, Jie Zheng, Benjamin Elsworth, Kaitlin H Wade, Valeriia Haberland, Denis Baird, Charles Laurin, Stephen Burgess, Jack Bowden, Ryan Langdon, Vanessa Y Tan, James Yarmolinsky, Hashem A Shihab, Nicholas J Timpson, David M Evans, Caroline Relton, Richard M Martin, George Davey Smith, Tom R Gaunt, and Philip C Haycock. The mr-base platform supports systematic causal inference across the human phenome. *eLife*, 7, 2018.

[28] Gibran Hemani, Kate Tilling, and George Davey Smith. Orienting the causal relationship between imprecisely measured traits using gwas summary data. *PLoS Genetics*, 13(12), 2017.

[29] Sang Hong Lee, Michael E Goddard, Naomi R Wray, and Peter M Visscher. A better coefficient of determination for genetic profile analysis. *Genetic Epidemiology*, 36(3), 2012.

[30] Yi Liu, Benjamin Elsworth, Pau Erola, Valeriia Haberland, Gibran Hemani, Matt Lyon, Jie Zheng, Oliver Lloyd, Marina Vabistsevits, and Tom R Gaunt. Epigraphdb: a database and data mining platform for health data science. *Bioinformatics*, 37(9):1304–1311, 2021.

8 Supplemental Note for Mind the gap: the population mismatch-sample size tradeoff of the same-population assumption in two sample Mendelian randomization

8.1 2SMR Analysis

For each 2SMR analysis, we identified genetic variants contained in both the exposure and outcome GWAS. We then LD clumped these overlapping variants, prioritizing variants based on their exposure study association p-value using an r^2 threshold of 0.01 and distance threshold of 500 kb. We selected the lead variant in each clump to be an instrumental variable if its p-value was less than either 10^{-5} or 5×10^{-8}. We then merged and harmonized exposure and outcome association estimates for the selected instruments.

To address horizontal pleiotropy, we included an overdispersion term in GRAPPLE, and used a Tukey loss function. To address sample overlap, we estimated the residual correlation between exposure and outcome effect estimates using the p-value thresholding approach described by Wang et al. [5]. However, we used a threshold of $p > 0.05$ instead of $p > 0.50$ as suggested by Wang et al. [5], as the latter tends to underestimate this shared correlation. We used the implementation of inverse variance weighted (IVW) regression in the `mr_ivw` function of the TwoSampleMR package [27].

We performed Steiger filtering on selected IVs for each study pair to remove variants more strongly correlated with the outcome than the exposure [28]. Steiger filtering can remove pleiotropic variants and reduce the risk of false positives due to reverse causation. In the Steiger filtering procedure, the proportions of exposure and outcome variance explained by each SNP are estimated and variants are removed if the proportion of outcome variance explained excedes the proportion of exposure variance explained. For continuous traits, the proportion of variance explained by variant G_j can be approximated as

$$R_j^2 \approx \frac{F_j}{F_j - 2 + n},$$

where n is the sample size of the GWAS study, $F_j = (\frac{\hat{\beta}_j}{\hat{se}(\hat{\beta}_j)})^2$, $\hat{\beta}_j$ is the estimated association of G_j with the trait, and $\hat{se}(\hat{\beta}_j)$ is the corresponding estimate of the standard error. For binary traits, we estimated the proportion of variance explained using the `get_r_from_lor()` function in the TwoSampleMR package, which implements an approximation provided by Lee et al. [29]. This function requires the estimated log odds ratio, case and control counts, the effect allele frequency of the variant, and the estimated prevalence of the binary trait in the study population. Prevalences for outcome studies were estimated using external literature, as detailed in Supplemental Data 4. Effect allele frequencies were not provided for four majority European ancestry outcome studies. For these studies, we approximated effect allele frequencies using the EUR superpopulation of 1000 Genomes.

8.2 Estimation with Measurement Error using Simulation Extrapolation (SIMEX)

To estimate the overall and population-pair specific shrinkage coefficients described in Methods Section 3.2.3, we performed regression of population-mismatched 2SMR estimates against their corresponding exact-matching reference estimates. To account for noise in the reference estimates, we used simulation extrapolation (SIMEX) [22]. In the SIMEX procedure, we add simulated noise to our real data and recompute the inverse-variance weighted regression coefficient. We perform this procedure for a range of noise level and use the trend in estimates to extrapolate the estimate we would have obtained if the reference estimates had been measured without noise.

Let $\hat{\gamma}_{1,mm}, \ldots, \hat{\gamma}_{J,mm}$ be the population-mismatched estimates used in the analysis. These serve as the outcome in the inverse-variance weighted regression. Let $\hat{\gamma}_{1,r}, \ldots, \hat{\gamma}_{J,r}$ be the corresponding reference (population-matched) estimates, which serve as the covariate. For a grid of values, λ, between 0 and 5 in intervals of 0.25, we sample new population-mismatched and reference estimates, $\hat{\gamma}_{j,mm,\lambda} \sim N(\hat{\gamma}_{j,mm}, \sigma^2_{j,mm})$ and $\hat{\gamma}_{j,r,\lambda} \sim N(\hat{\gamma}_{j,r}, (1+\lambda)\sigma^2_{j,r})$, where $\sigma_{j,r}$ and $\sigma_{j,mm}$ are the estimated standard errors of $\hat{\gamma}_{j,r}$ and $\hat{\gamma}_{j,mm}$ respectively. For each value of λ, we compute $\hat{\beta}_\lambda$, the coefficient in the inverse-variance weighted regression. The condition of no measurement error in $\hat{\gamma}_{j,r}$ corresponds to the condition that $\lambda = -1$, which is not observable. We therefore extrapolate this value by fitting a function to approximate $\hat{\beta}_\lambda$ as a function of λ and then plugging in $\lambda = -1$. We found that the a linear function fit the data well for our overall shrinkage coefficients (Figure S3), as well as for population pair-specific shrinkage coefficients with many estimate groups available (such as using a UK Biobank exposure for a Biobank Japan target). For population pairs with fewer estimate groups available or less precise regression results, observed data were generally too noisy to rule out a linear extrapolation function. To estimate a standard error for $\hat{\beta}_{-1}$, we repeat this procedure several times resampling the observations that contribute to the regression estimate. The standard error of $\hat{\beta}_{-1}$ is estimated as the standard deviation of the bootstrap estimates.

8.3 Population pair-specific analyses at a more stringent instrument selection threshold

We used SIMEX to estimate population pair-specific shrinkage coefficients using the instrument selection p-value threshold of 5×10^{-8} in addition to the more lenient 10^{-5} threshold. However, at the stricter threshold, no population pair with an

African ancestry GWAS used as the outcome had more than 8 population-mismatched estimates that could be compared to a reference, leaving only 23 population pairs in which SIMEX estimates were computed. The slope of inverse variance weighted regression of estimated shrinkage coefficients vs F_{st} is -2.51, with a standard error of 0.70 ($p = 0.003$). This is similar to the slope of -2.40 obtained when using more lenient threshold of 1×10^{-5}.

SIMEX-estimated shrinkage coefficients and bootstrap standard errors using both instrument selection thresholds are shown in Supplementary Figure S2. Supplementary Figure Figure S4 shows reference vs population-mismatched estimates for each population pair. At both thresholds, most of these slopes are much lower than 1, aligning with our predictions and the overall trends observed in the main manuscript. In some cases, the SIMEX-estimated shrinkage coefficient is negative, usually when using GWAS sampled from Africans in the Pan-UKB project as the exposure or target population. However, 95% confidence intervals of these coefficients always included zero.

There are some differences in shrinkage coefficients between instrument selection thresholds. For example, the FinnGen exposure - UK Biobank outcome is lower at the more lenient threshold (0.53 for 1×10^{-5} vs. 0.65 for 5×10^{-8}), while the UK Biobank exposure - Biobank Japan outcome estimate is larger at the more lenient threshold (0.44 for 1×10^{-5} vs. 0.26 for 5×10^{-8}).

8.4 Supplementary Figures

Trait	Matching Outcome	Standard unit	Number of studies
LDL cholesterol	Coronary artery disease	mmol/L	5
Body Mass Index	Type 2 diabetes	kg/m^2	6
Hemoglobin A1c	Coronary artery disease	mmol/mol	4
Systolic Blood Pressure	Ischemic stroke	mmHg	5
Eosinophil count	Rheumatoid arthritis	10^9 cells/L	4
HDL cholesterol	Breast cancer	mmol/L	6
Lymphocyte count	Schizophrenia	10^9 cells/L	6
Height	Atrial fibrillation	cm	7
Vitamin D	Multiple sclerosis	nmol/L	3
Weight	Myocardial infarction	kg	6
C-reactive protein	Alzheimer's disease	mg/L	4
Bone mineral density	Osteoporosis	z-score	3
Urea	Gout	mmol/L	6
Basophil counts	Atopic dermatitis	10^9 cells/L	6
Hip circumference	Crohn's disease	cm	3
Mean corpuscular hemoglobin	Total bilirubin	pg	3
Mean corpuscular hemoglobin concentration	Iron deficiency anemia	g/dL	4
Monocyte count	Asthma	10^9 cells/L	6
Neutrophil count	COPD	10^9 cells/L	6
Platelet count	Hypertension	10^9 cells/L	6
RBC count	Total bilirubin	10^{12} cells/L	6
Triglycerides	Hyperlipidemia	mmol/L	5
Uric acid	Gout	mg/dL	6
Waist-hip ratio	Type 2 diabetes	Unitless	3
Waist circumference	NAFLD/MASLD	cm	3
WBC count	Coronary artery disease	10^9 cells/L	6

Table S1: A summary of the 26 exposure traits we included in our analysis, and the standard unit used for rescaling 2SMR estimates involving the exposure. Each of these exposure traits was linked to a matching outcome trait that we expected it to have a likely non-null effect on, either through literature or via EpiGraphDB [30]. More details on the specific studies for each of these exposure traits is located in Supplemental Data 2.

Outcome Trait	Number of Studies
Atrial Fibrillation	5
Alzheimer's	2
Anemia	4
Asthma	4
Atopic Dermatitis	4
Bilirubin	4
Breast cancer	4
Coronary Heart Disease	5
COPD	6
Crohn's disease	3
Gout	6
Hyperlipidemia	3
Hypertension	5
Myocardial Infarction	5
Multiple Sclerosis	3
NAFLD/MASLD	2
Osteoporosis	3
Rheumatoid arthritis	5
Schizophrenia	4
Stroke	5
Type 2 diabetes	2

Table S2: A summary of the 21 outcome traits we included in our analysis, including its standard unit for rescaling and the number of studies for each trait. More details on the specific studies for each of these outcome traits is located in Supplemental Data 3.

Subpopulation	Scott et al 2016 Equivalent
UK Biobank	Great Britain
Biobank Japan	Japanese
FinnGen	Finnish
Taiwan Biobank	Chinese (South)
Mixed (majority Hispanic) mega analysis (Wojcik)	No equivalent
Pan-UKB South Asians	Persia and Pakistan
Pan-UKB Africans	Yoruba (Nigeria)

Table S3: Closest equivalents of subpopulations in our 2SMR survey to the populations in Scott et al 2016.

Total Number of Mismatching Estimates with Reference Estimates	1,951
Mismatching Estimates with Non-Significant Concordance z-scores	1,777
Mismatching Estimates with Significant Concordance z-scores, Same Sign as Reference, Smaller Magnitude	101
Mismatching Estimates with Significant Concordance z-scores, Same Sign as Reference, Larger Magnitude	7
Mismatching Estimates with Significant Concordance z-scores, Opposite Sign as Reference	66
Mismatching Estimates with Significant Concordance z-scores, Opposite Sign as Reference Reference and Mismatch 2SMR p-values < 0.05	15
SIMEX-Estimated Overall Shrinkage Coefficient, 1000 Bootstrap Resamples	0.453 (SE 0.018)

Table S4: A summary of how 2SMR estimates with mismatching exposures compare to their corresponding reference 2SMR estimates. 2SMR estimates were computed using the inverse-variance weighted (IVW) regression method at an instrument selection threshold of 5×10^{-8}.

Trait Pair	IV Selection Threshold	Exposure Population	Target Population	Mismatching estimate (p-value)	Reference estimate (p-value)	Concordance z-score
Body Mass Index to Breast Cancer	5×10^{-8}	Biobank Japan	UK Biobank	-0.074 ($p < 0.0001$)	-0.006 ($p = 0.44$)	-3.36
HDL cholesterol to Type 2 diabetes	1×10^{-5}	UK Biobank	Biobank Japan	-0.47 ($p < 0.0001$)	-0.04 ($p = 0.55$)	-4.48
HDL cholesterol to Type 2 diabetes	5×10^{-8}	UK Biobank)	Biobank Japan	-0.41 ($p < 0.0001$)	-0.03 ($p = 0.68$)	-3.40
Height to Atrial Fibrillation	5×10^{-8}	FinnGen	UK Biobank	0.064 ($p < 0.0001$)	0.039 ($p < 0.0001$)	3.35
Weight to Breast Cancer	5×10^{-8}	Taiwan Biobank	UK Biobank	-0.023 ($p = 0.0009$)	-0.001 ($p = 0.60$)	-2.94
Weight to Type 2 diabetes	5×10^{-8}	Taiwan Biobank	Biobank Japan	0.054 ($p < 0.0001$)	0.012 ($p = 0.20$)	2.87

Table S5: The concordance z-scores and estimate effect sizes of mismatching 2SMR estimates that are larger in magnitude than their corresponding references in the same direction, and with significant concordance z-scores at an FDR of 5%.

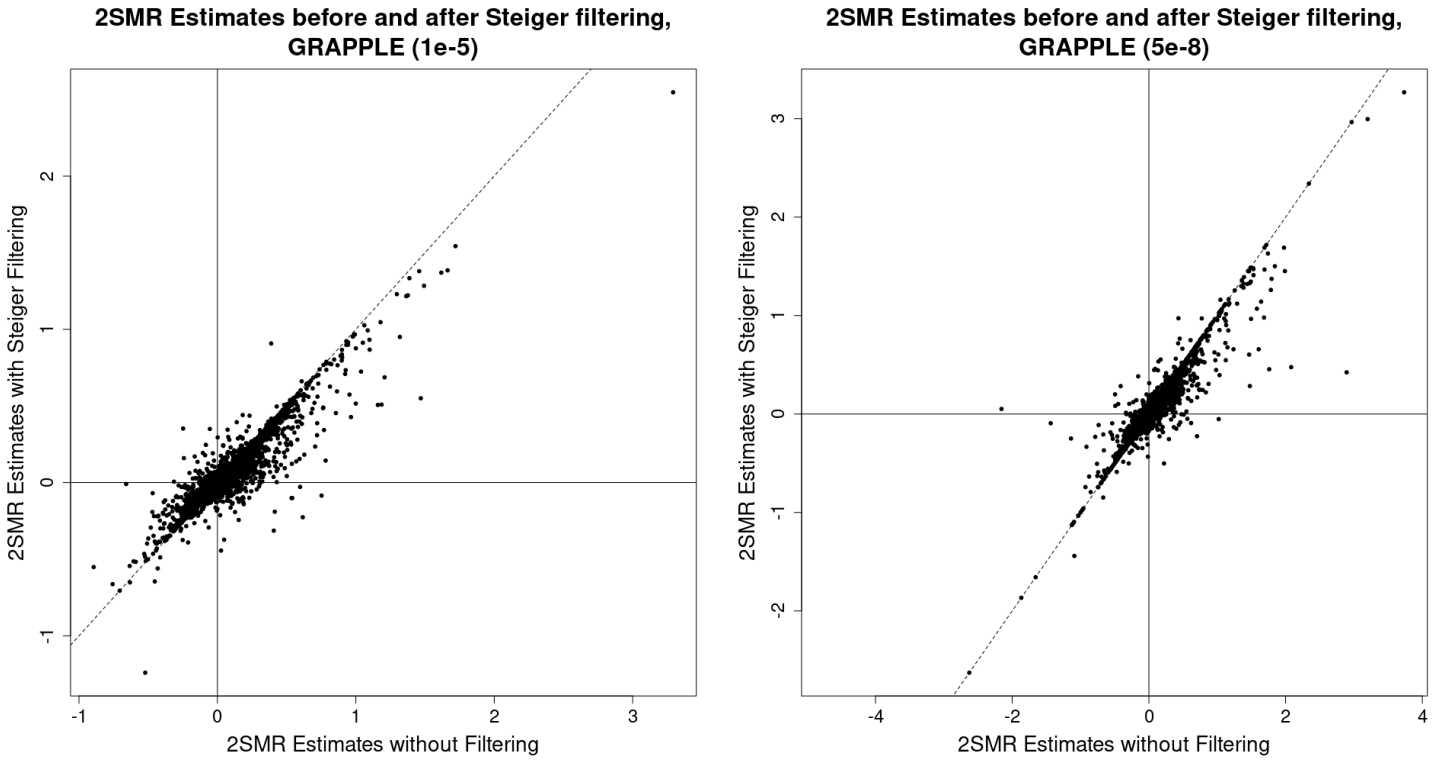


Figure S1: A comparison of 2SMR estimates estimated using GRAPPLE before and after Steiger filtering at two instrument selection thresholds. The dashed line represents the $y = x$ line. All estimates were not yet rescaled to the respective standard units of their exposure.

A) GRAPPLE, 1×10^{-5}

Target → Exposure ↓	UK Biobank	FinnGen	Biobank Japan	Taiwan Biobank	Hispanic mega-analysis	SAS (UK)	AFR (UK)
UK Biobank		0.67 (0.02) [40]	0.44 (0.03) [260]	0.15 (0.09) [9]	0.68 (0.17) [12]	0.11 (0.05) [40]	0.02 (0.13) [8]
FinnGen	0.53 (0.02) [36]		0.50 (0.05) [26]				
Biobank Japan	0.25 (0.02) [360]	0.32 (0.02) [40]			0.23 (0.14) [12]	0.14 (0.08) [40]	-0.32 (0.25) [8]
Taiwan Biobank	0.40 (0.02) [162]	0.34 (0.04) [20]	0.77 (0.04) [91]				
Hispanic mega-analysis	0.21 (0.02) [108]	0.15 (0.02) [40]	0.36 (0.04) [78]			0.30 (0.24) [8]	
SAS (UK)	0.06 (0.02) [180]		0.06 (0.03) [130]				-0.34 (0.19) [8]
AFR (UK)	-0.01 (0.01) [144]		-0.01 (0.02) [104]			0.07 (0.03) [32]	

B) GRAPPLE, 5×10^{-8}

Target → Exposure ↓	UK Biobank	FinnGen	Biobank Japan	Taiwan Biobank	Hispanic mega-analysis	SAS (UK)	AFR (UK)
UK Biobank		0.70 (0.03) [40]	0.26 (0.03) [260]	0.14 (0.09) [9]	0.40 (0.11) [12]	0.04 (0.03) [36]	
FinnGen	0.65 (0.03) [36]		0.53 (0.05) [26]				
Biobank Japan	0.31 (0.02) [360]	0.39 (0.03) [40]			0.29 (0.14) [12]	0.14 (0.05) [36]	
Taiwan Biobank	0.56 (0.03) [162]	0.38 (0.06) [20]	0.67 (0.05) [91]				
Hispanic mega-analysis	0.52 (0.03) [108]	0.47 (0.04) [40]	0.57 (0.05) [78]			0.12 (0.14) [8]	
SAS (UK)	0.02 (0.10) [124]		0.04 (0.05) [87]				
AFR (UK)	0.01 (0.02) [108]		-0.03 (0.02) [78]			0.03 (0.02) [24]	

Figure S2: Population pair-specific SIMEX-estimated shrinkage coefficients. Bootstrap standard error is given in round parentheses. The number of estimates contributing to each coefficient is given in square parentheses. No estimate was computed when fewer than eight estimates for a population pair were available.

Trait Pair	Significance of Meta-analysis and BBJ Exposures	Populations Used in Replication	Replication Status
Weight to Coronary Heart Disease	Meta-analysis Only	FinnGen	Replicated
Hemoglobin A1c to Coronary Heart Disease	Meta-analysis Only		Not Tested
Monocyte count to Bilirubin	Meta-analysis Only	AFR (Pan-UKB), SAS (Pan-UKB)	Not Replicated
RBC count to Bilirubin	Meta-analysis Only	AFR (Pan-UKB), SAS (Pan-UKB)	Replicated
Hemoglobin A1c to Bilirubin	Meta-analysis Only		Not Tested
Height to COPD	Meta-analysis Only	FinnGen	Not Replicated
RBC count to Myocardial Infarction	Meta-analysis Only		Not Tested
Height to Myocardial Infarction	Meta-analysis Only	FinnGen	Replicated
Weight to Myocardial Infarction	Meta-analysis Only	FinnGen	Replicated
Hemoglobin A1c to Myocardial Infarction	Meta-analysis Only		Not Tested
Weight to Stroke	Meta-analysis Only	FinnGen	Replicated
C-reactive protein to Type 2 diabetes	Meta-analysis Only	Majority Hispanic mega-analysis	Replicated
HDL cholesterol to Type 2 diabetes	Meta-analysis Only	Majority Hispanic mega-analysis	Replicated
Triglycerides to Type 2 diabetes	Meta-analysis Only	Majority Hispanic mega-analysis	Replicated
HDL cholesterol to Bilirubin	BBJ Only		Not Tested
Height to Asthma	BBJ Only	FinnGen	Replicated
Triglycerides to COPD	BBJ Only		Not Tested
Weight to Atrial Fibrillation	BBJ Only	FinnGen	Replicated
Height to Type 2 diabetes	BBJ Only	Majority Hispanic mega-analysis, FinnGen	Replicated

Table S6: Replication analysis results in non-UK and non-Japanese populations for 19 causal effect estimates that differed in significance ($p < 0.05/195$) when the exposure was measured either only in Biobank Japan or in a meta analysis of UK Biobank and Biobank Japan using instrument selection threshold 10^{-5}. An effect was considered replicated if it had a non-UK, non-Japanese 2SMR estimate in the same direction as the significant estimate, with a p-value below a Bonferroni-corrected threshold (0.05/13).

Trait Pair	Significance of Meta analysis and BBJ Exposures	Populations Used in Replication	Replication Status
Height to Coronary Heart Disease	Meta analysis Only	FinnGen	Replicated
Hemoglobin A1c to Coronary Heart Disease	Meta analysis Only		Not Tested
Monocyte count to Bilirubin	Meta analysis Only	AFR (Pan UKB), SAS (Pan UKB)	Not Replicated
RBC count to Bilirubin	Meta analysis Only	AFR (Pan UKB), SAS (Pan UKB)	Not Replicated
Hemoglobin A1c to Bilirubin	Meta analysis Only		Not Tested
Height to Asthma	Meta analysis Only	FinnGen	Replicated
Height to COPD	Meta analysis Only	FinnGen	Not Replicated
Height to Myocardial Infarction	Meta analysis Only	FinnGen	Replicated
Weight to Myocardial Infarction	Meta analysis Only	FinnGen	Replicated
Hemoglobin A1c to Myocardial Infarction	Meta analysis Only		Not Tested
Platelet count to Type 2 diabetes	Meta analysis Only	Majority Hispanic mega analysis	Not Replicated
Weight to Type 2 diabetes	Meta analysis Only	FinnGen	Replicated
HDL cholesterol to Type 2 diabetes	Meta analysis Only	Majority Hispanic mega analysis	Not Replicated
Hemoglobin A1c to Type 2 diabetes	Meta analysis Only	Majority Hispanic mega analysis	Not Replicated
Triglycerides to COPD	BBJ Only		Not Tested
Weight to Atrial Fibrillation	BBJ Only	FinnGen	Replicated

Table S7: Replication analysis results in non-UK and non-Japanese populations for 19 causal effect estimates that differed in significance ($p < 0.05/195$) when the exposure was measured either only in Biobank Japan or in a meta analysis of UK Biobank and Biobank Japan using instrument selection threshold 5×10^{-8}. An effect was considered replicated if it had a non-UK, non-Japanese 2SMR estimate in the same direction as the significant estimate, with a p-value below a Bonferroni-corrected threshold (0.05/12).

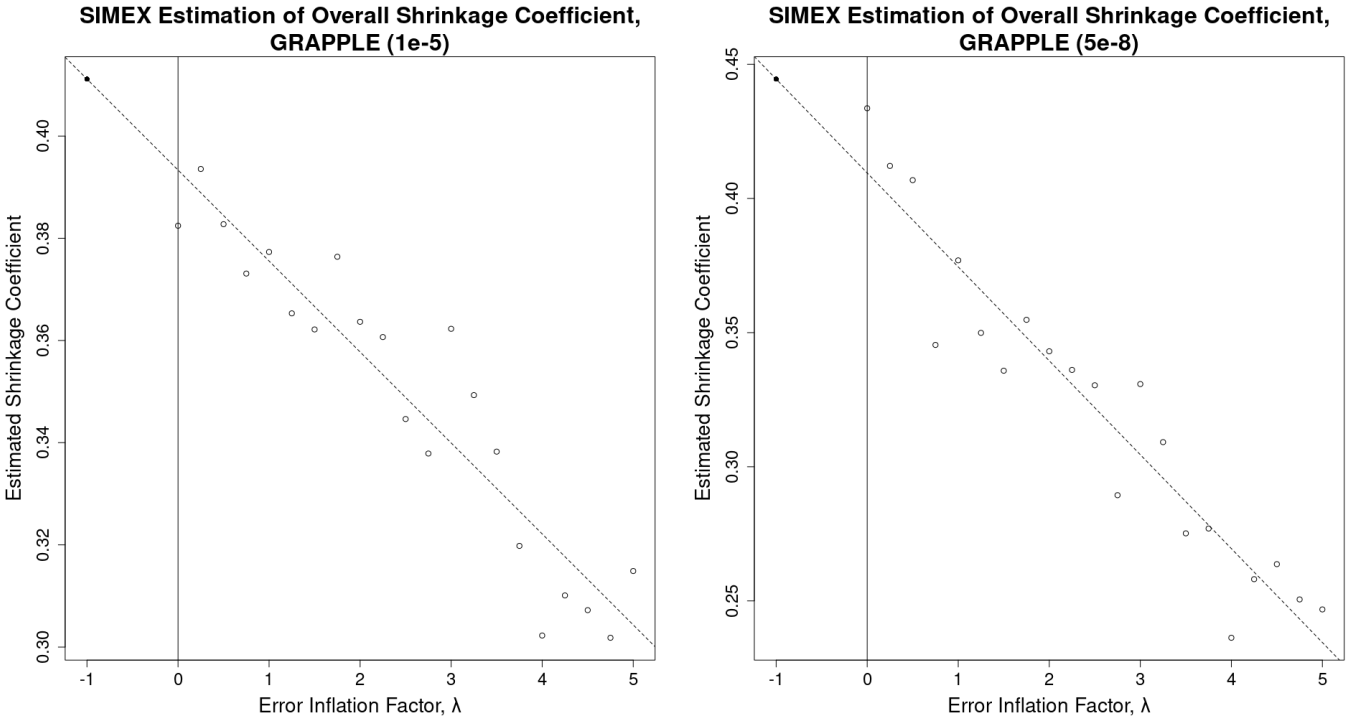
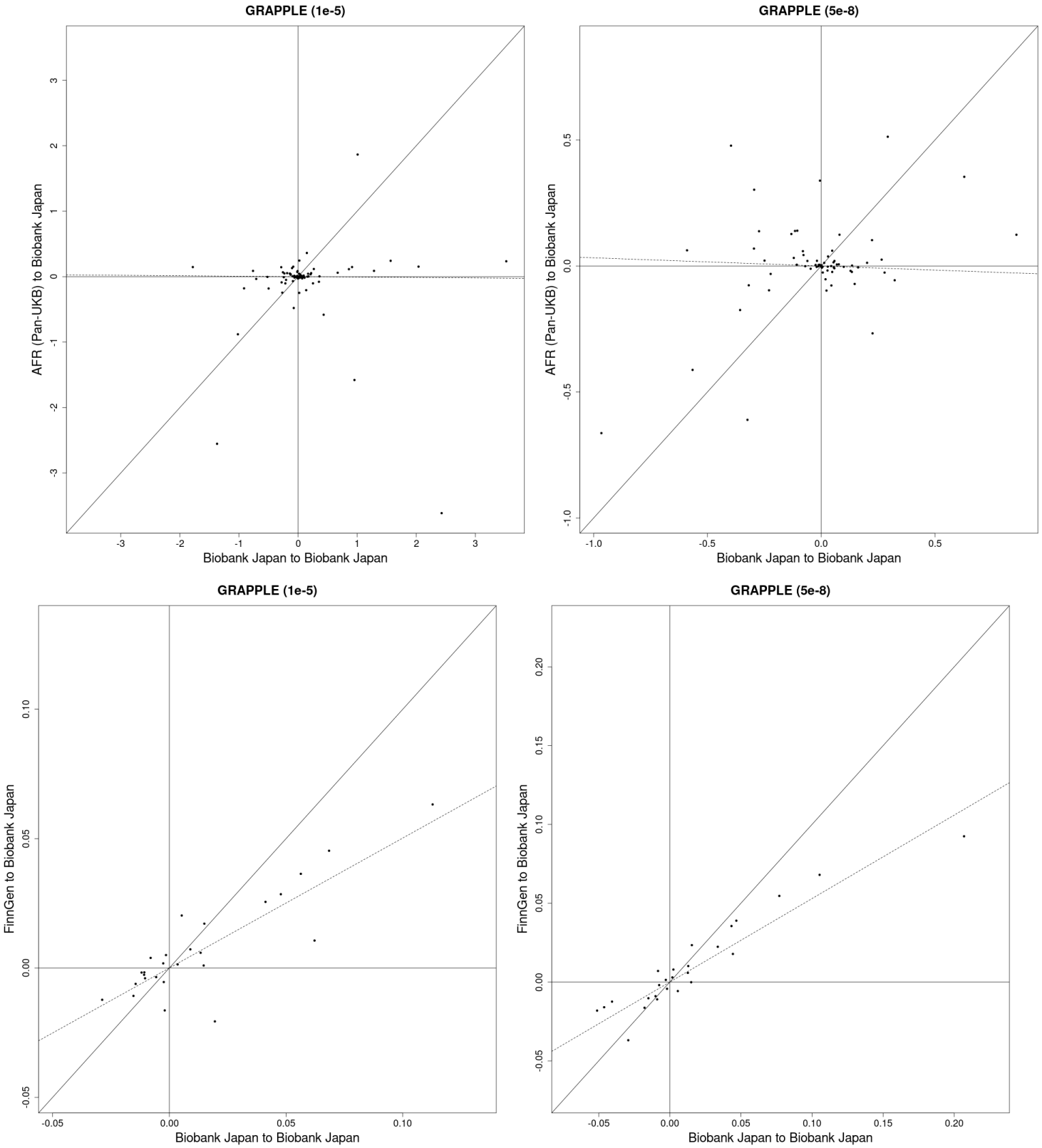
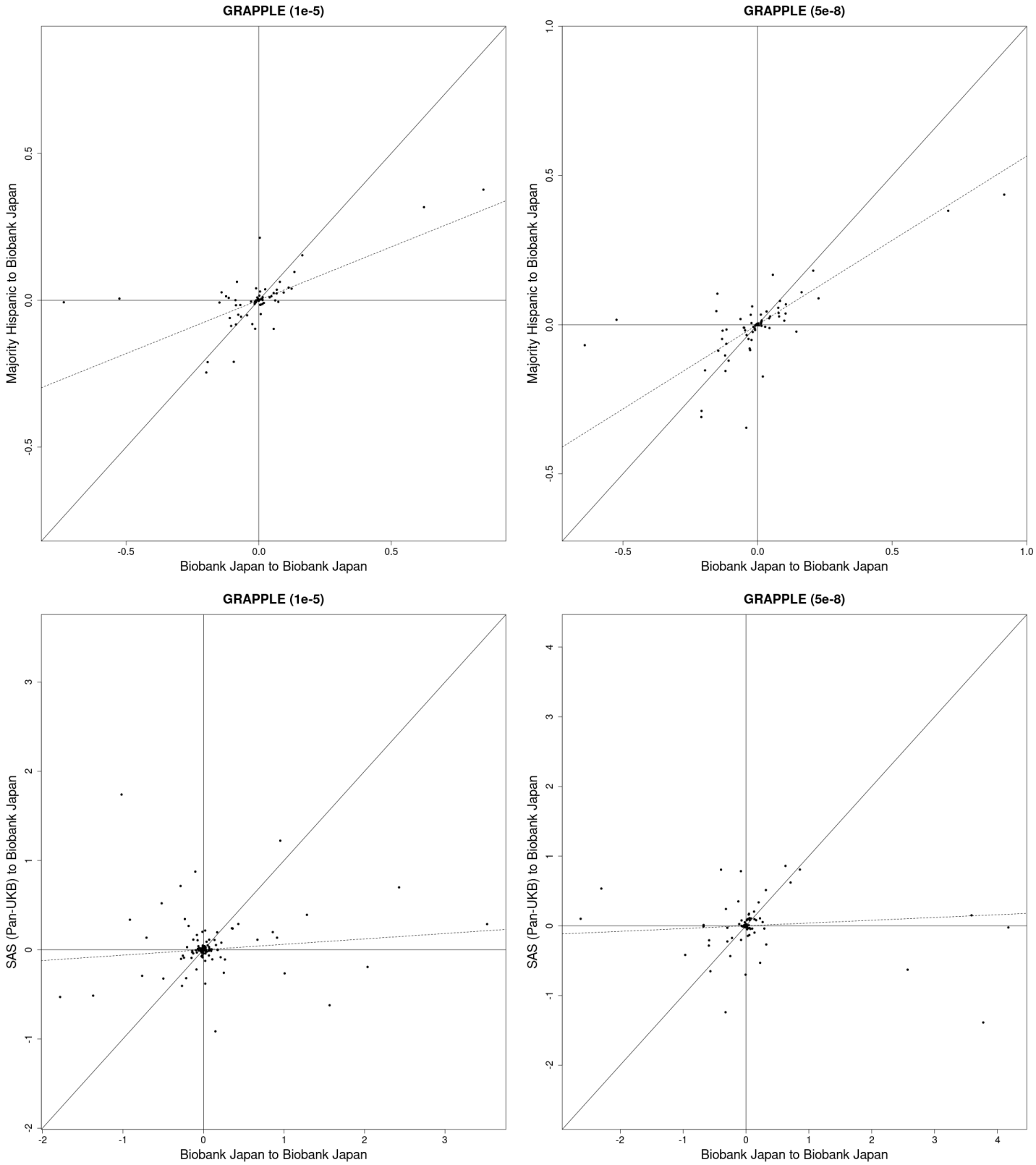
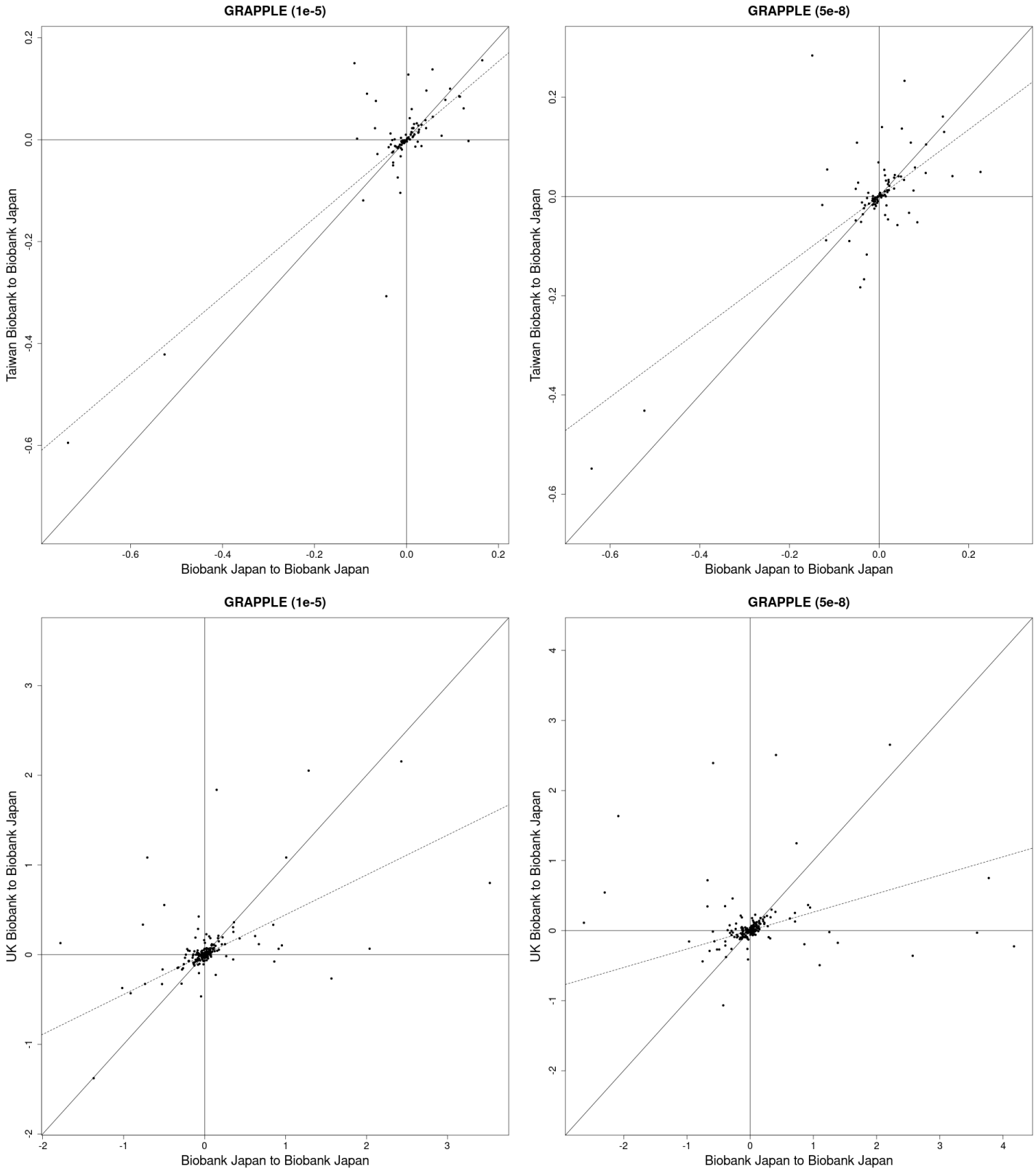
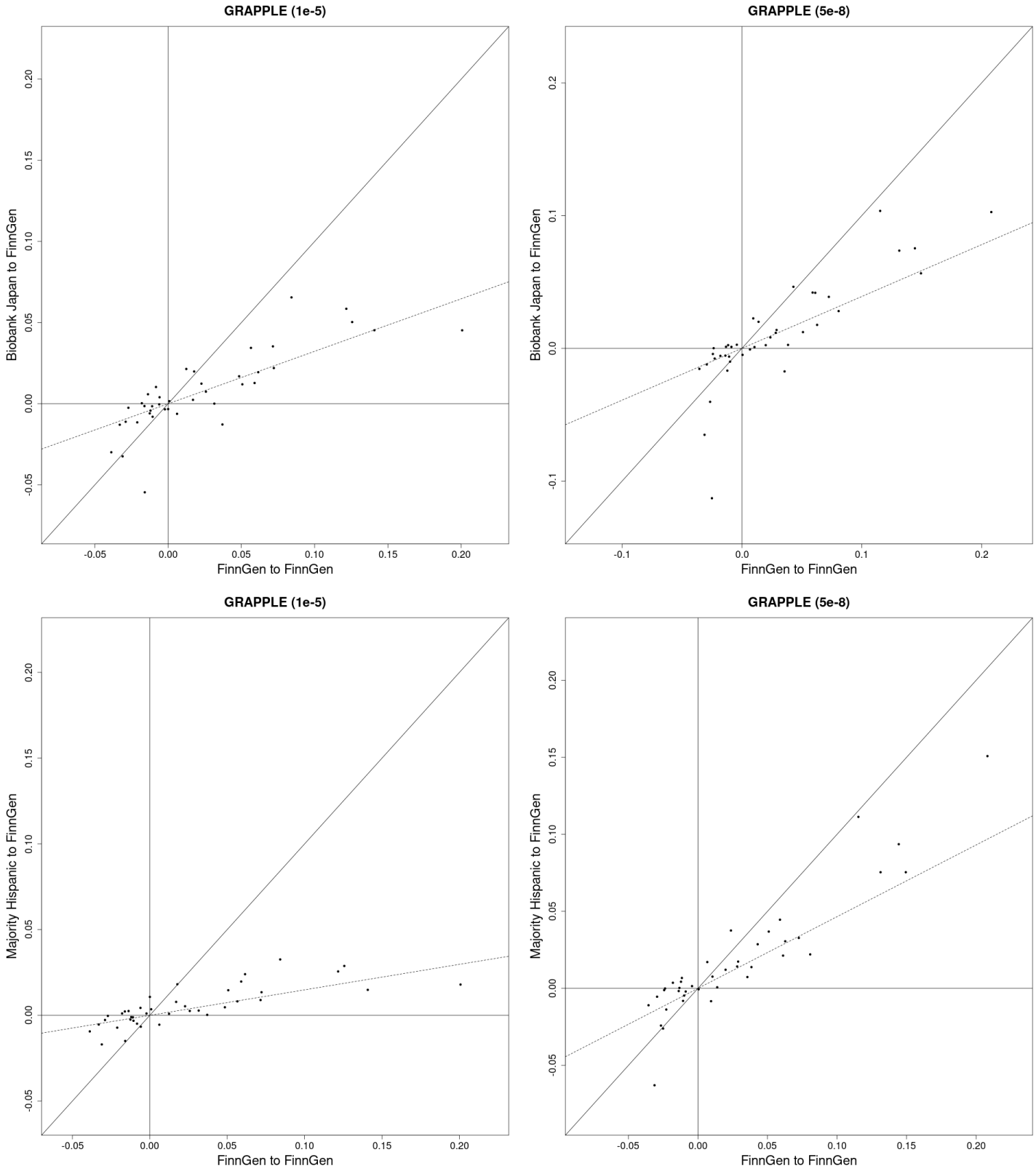


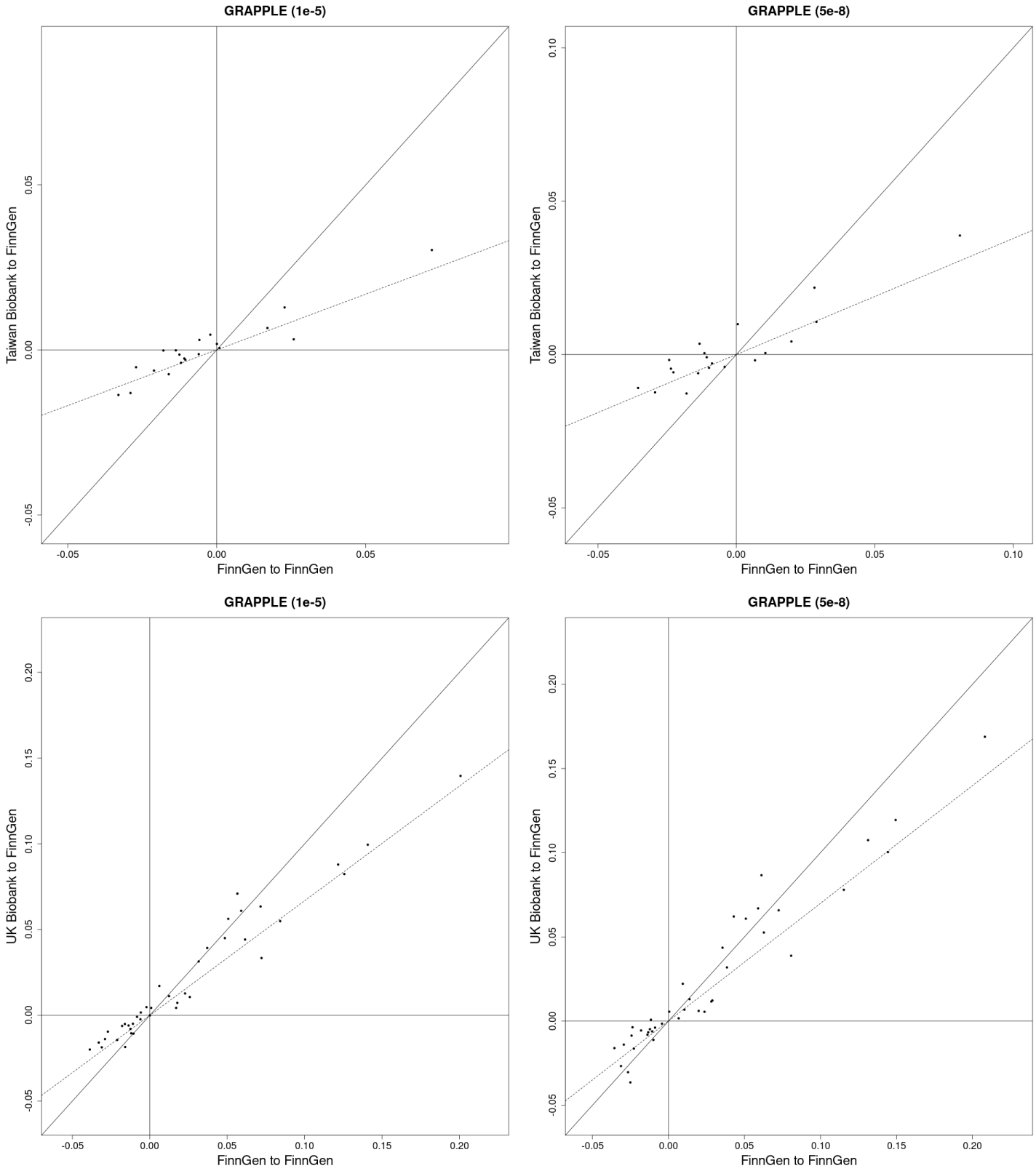
Figure S3: Estimated overall shrinkage coefficients of 2SMR vs additional measurement error (SIMEX λ) for two instrument selection thresholds. Estimated shrinkage coefficients with added noise are shown as open points. The SIMEX-estimated "noise-free" coefficient at $\lambda = -1$ is extrapolated using the noisier estimates, and is shown as a solid point. The slope and intercept of the linear regression used for extrapolation is shown as a dashed line.

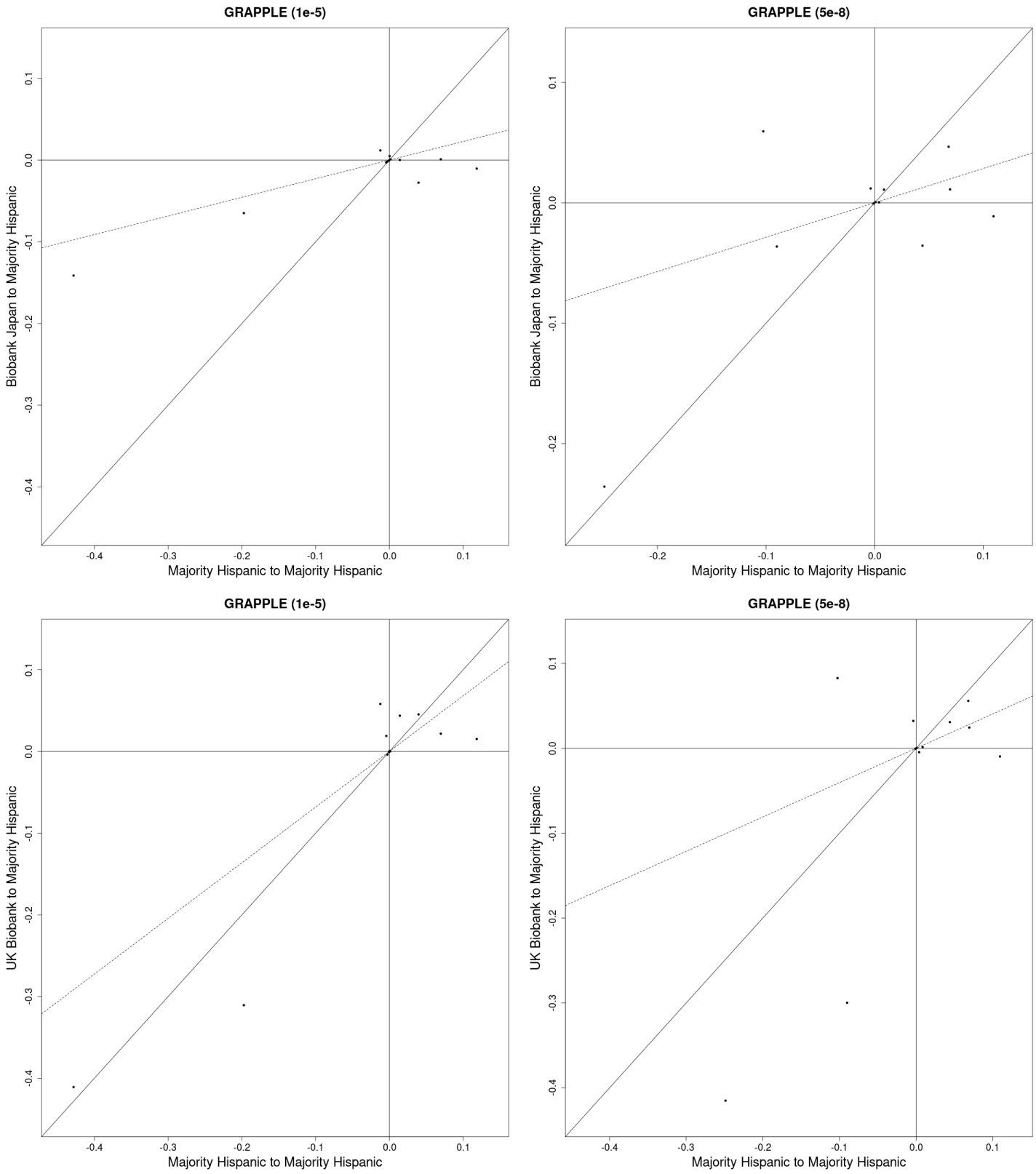

GRAPPLE (1e-5)
AFR (Pan-UKB) to Biobank Japan
Biobank Japan to Biobank Japan
GRAPPLE (5e-8)
AFR (Pan-UKB) to Biobank Japan
Biobank Japan to Biobank Japan
GRAPPLE (1e-5)
FinnGen to Biobank Japan
Biobank Japan to Biobank Japan
GRAPPLE (5e-8)
FinnGen to Biobank Japan
Biobank Japan to Biobank Japan

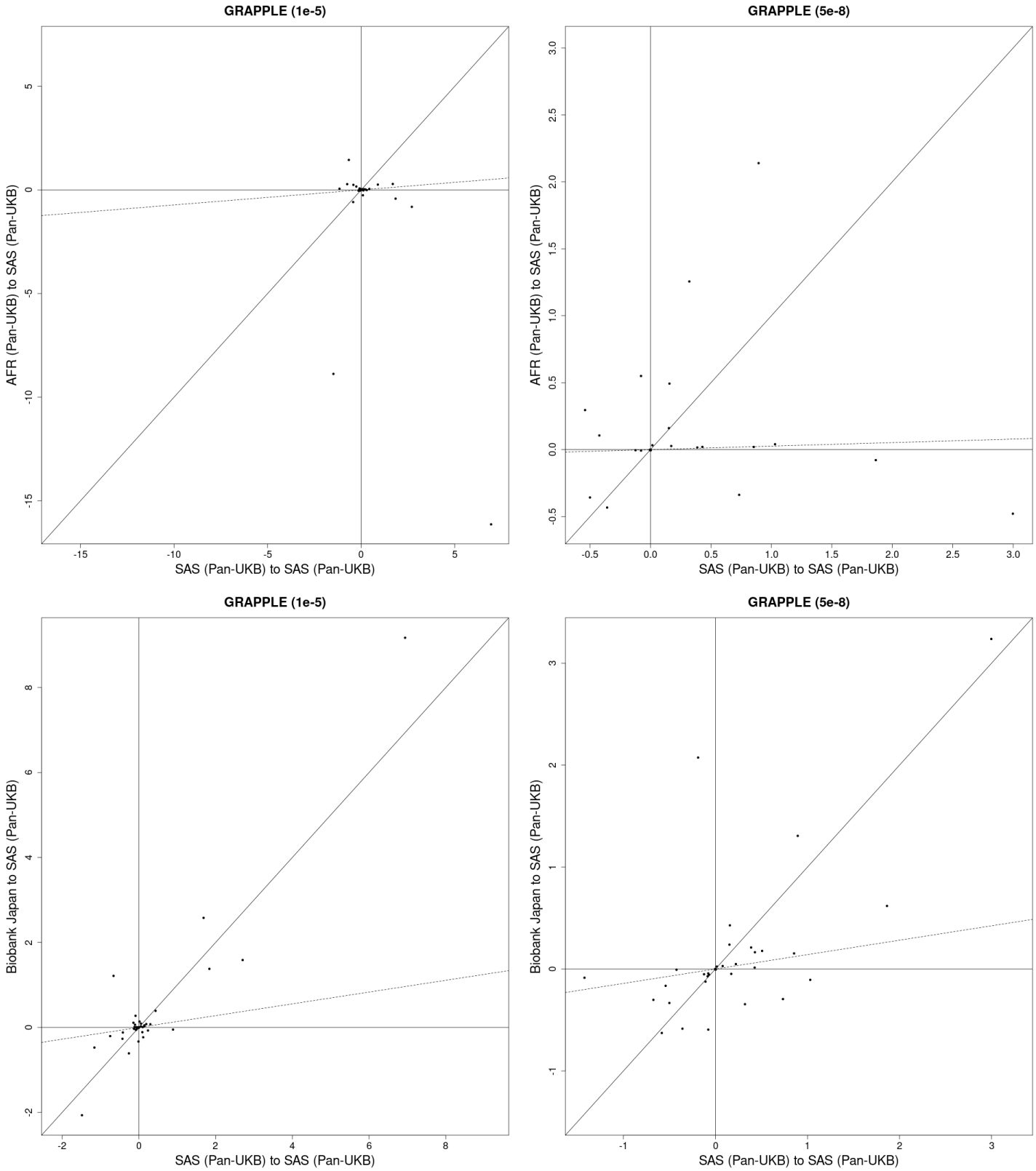

GRAPPLE (1e-5)
Majority Hispanic to Biobank Japan
Biobank Japan to Biobank Japan
GRAPPLE (5e-8)
Majority Hispanic to Biobank Japan
Biobank Japan to Biobank Japan
GRAPPLE (1e-5)
SAS (Pan-UKB) to Biobank Japan
Biobank Japan to Biobank Japan
GRAPPLE (5e-8)
SAS (Pan-UKB) to Biobank Japan
Biobank Japan to Biobank Japan

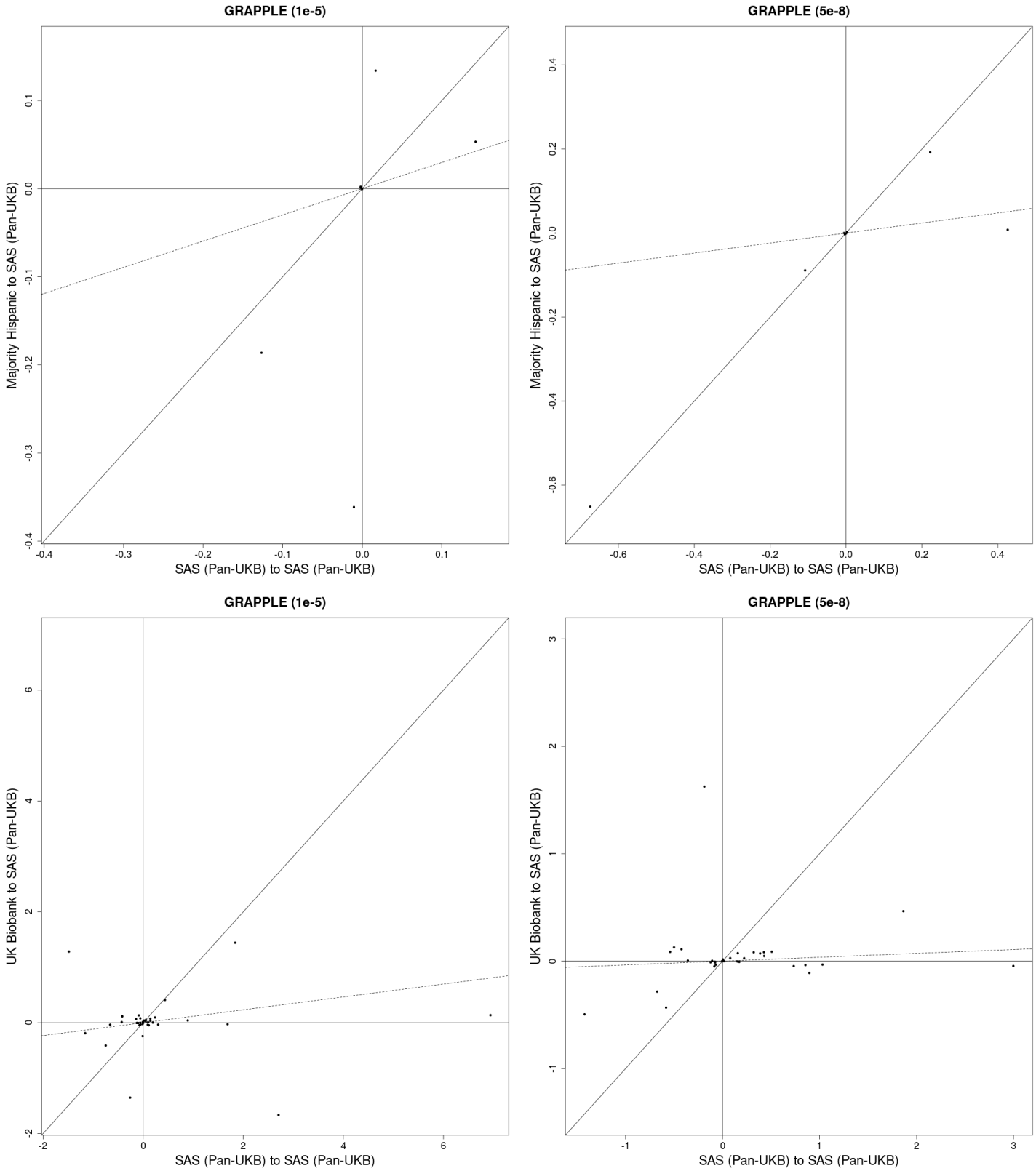

GRAPPLE (1e-5)
Taiwan Biobank to Biobank Japan
Biobank Japan to Biobank Japan
GRAPPLE (5e-8)
Taiwan Biobank to Biobank Japan
Biobank Japan to Biobank Japan
GRAPPLE (1e-5)
UK Biobank to Biobank Japan
Biobank Japan to Biobank Japan
GRAPPLE (5e-8)
UK Biobank to Biobank Japan
Biobank Japan to Biobank Japan

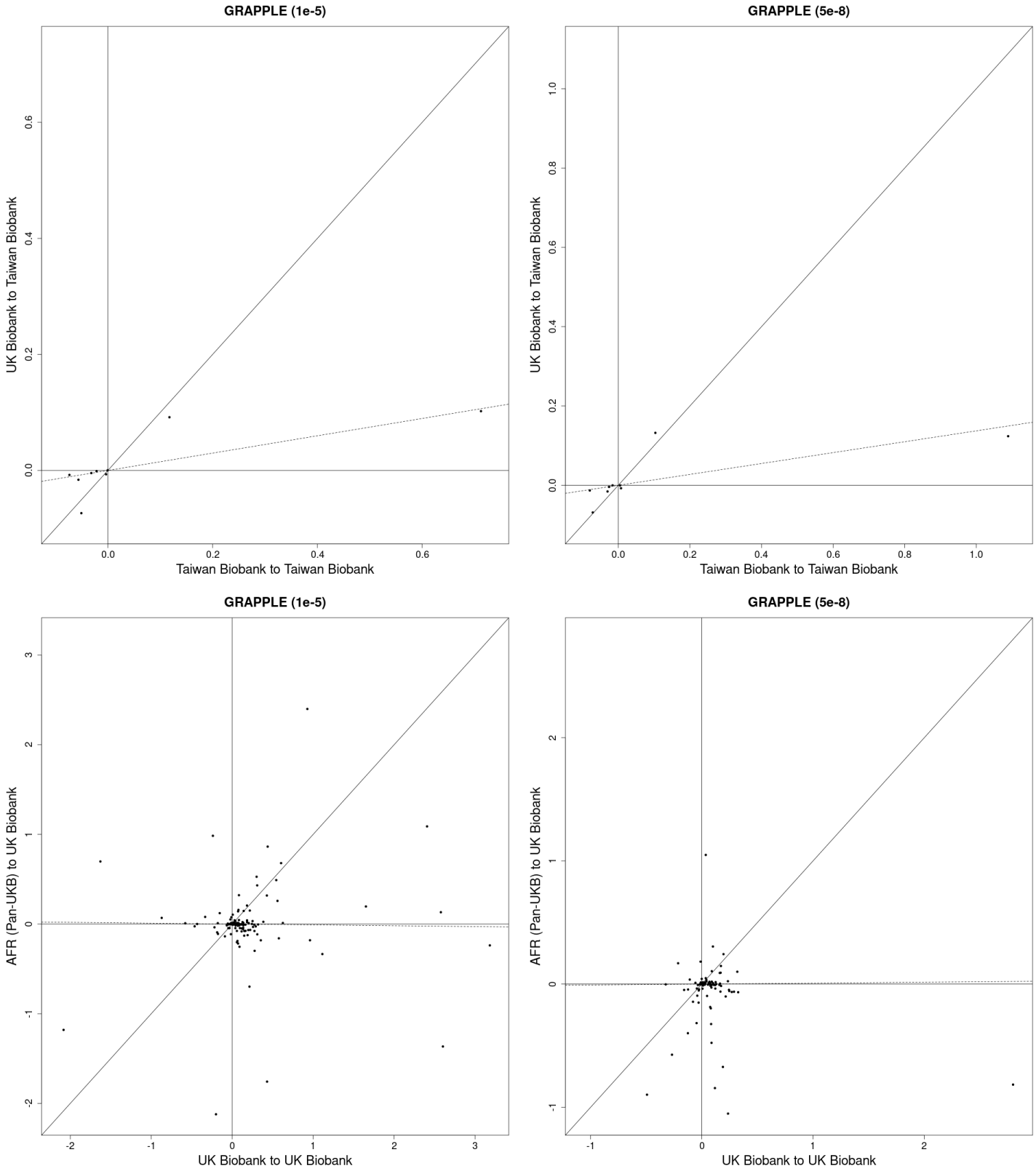

GRAPPLE (1e-5)
GRAPPLE (5e-8)
GRAPPLE (1e-5)
GRAPPLE (5e-8)
Biobank Japan to FinnGen
Majority Hispanic to FinnGen
FinnGen to FinnGen
-0.05
0.00
0.05
0.10
0.15
0.20
-0.1
0.0
0.1
0.2

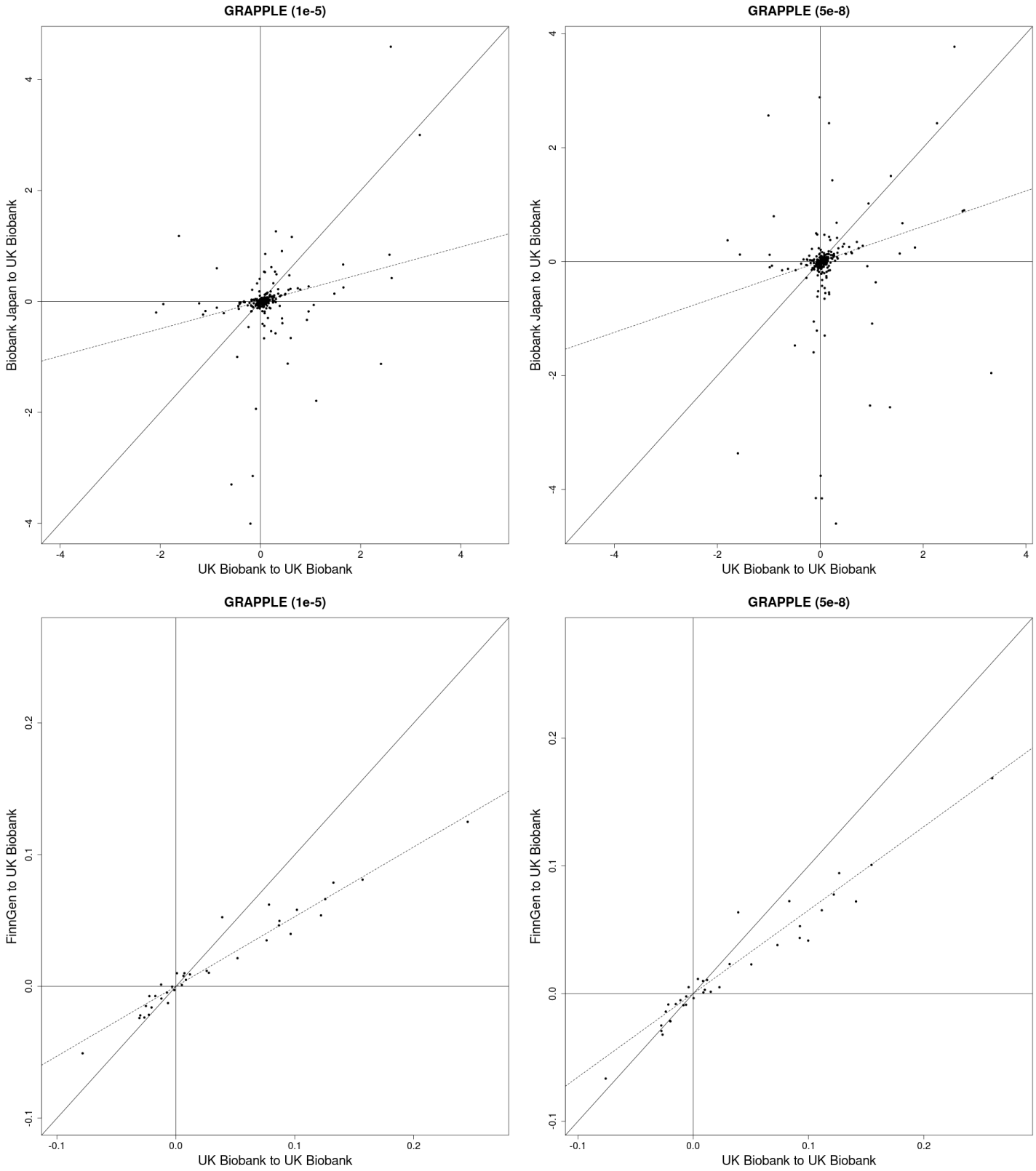

GRAPPLE (1e-5)
GRAPPLE (5e-8)
GRAPPLE (1e-5)
GRAPPLE (5e-8)
Taiwan Biobank to FinnGen
UK Biobank to FinnGen
FinnGen to FinnGen
-0.05
0.00
0.05
0.10
0.15
0.20

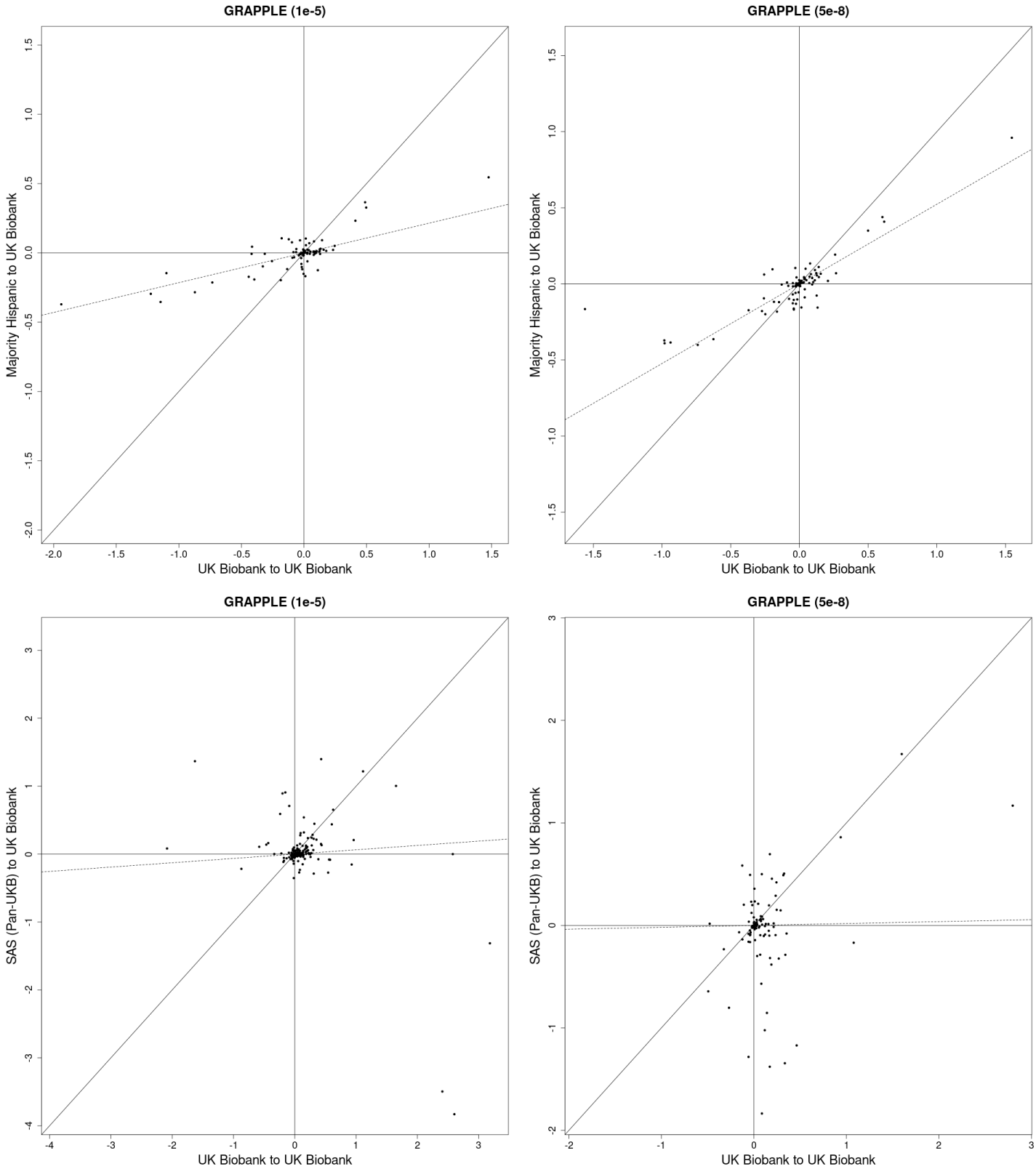

GRAPPLE (1e-5)
Biobank Japan to Majority Hispanic
Majority Hispanic to Majority Hispanic
GRAPPLE (5e-8)
Biobank Japan to Majority Hispanic
Majority Hispanic to Majority Hispanic
GRAPPLE (1e-5)
UK Biobank to Majority Hispanic
Majority Hispanic to Majority Hispanic
GRAPPLE (5e-8)
UK Biobank to Majority Hispanic
Majority Hispanic to Majority Hispanic

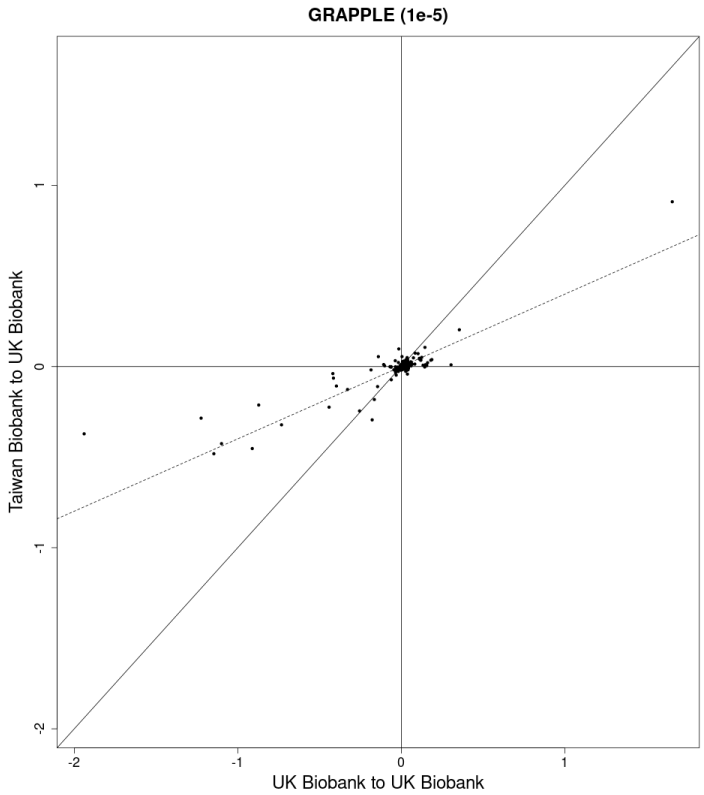

GRAPPLE (1e-5)
AFR (Pan-UKB) to SAS (Pan-UKB)
SAS (Pan-UKB) to SAS (Pan-UKB)
GRAPPLE (5e-8)
AFR (Pan-UKB) to SAS (Pan-UKB)
SAS (Pan-UKB) to SAS (Pan-UKB)
GRAPPLE (1e-5)
Biobank Japan to SAS (Pan-UKB)
SAS (Pan-UKB) to SAS (Pan-UKB)
GRAPPLE (5e-8)
Biobank Japan to SAS (Pan-UKB)
SAS (Pan-UKB) to SAS (Pan-UKB)

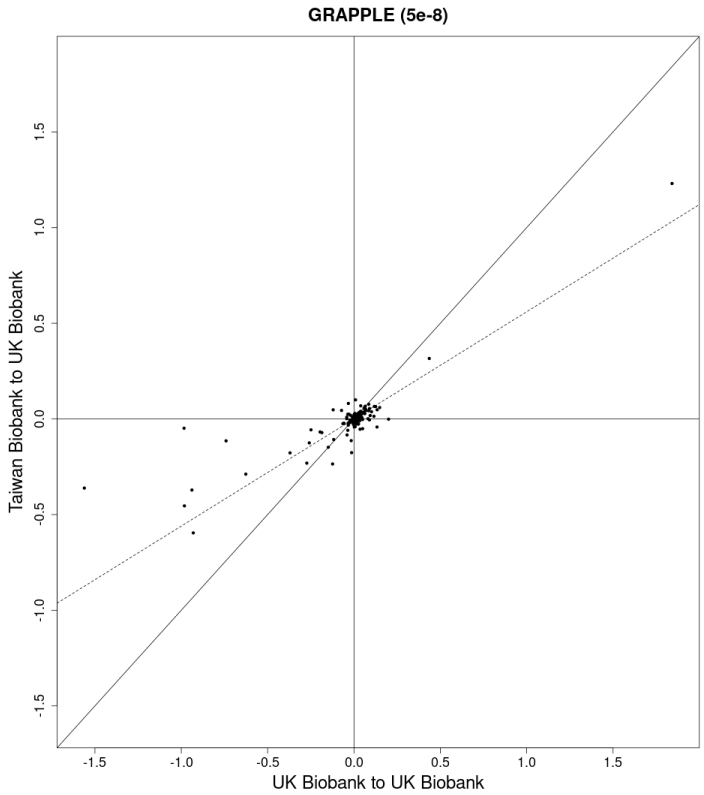

GRAPPLE (1e-5)
Majority Hispanic to SAS (Pan-UKB)
SAS (Pan-UKB) to SAS (Pan-UKB)
GRAPPLE (5e-8)
Majority Hispanic to SAS (Pan-UKB)
SAS (Pan-UKB) to SAS (Pan-UKB)
GRAPPLE (1e-5)
UK Biobank to SAS (Pan-UKB)
SAS (Pan-UKB) to SAS (Pan-UKB)
GRAPPLE (5e-8)
UK Biobank to SAS (Pan-UKB)
SAS (Pan-UKB) to SAS (Pan-UKB)

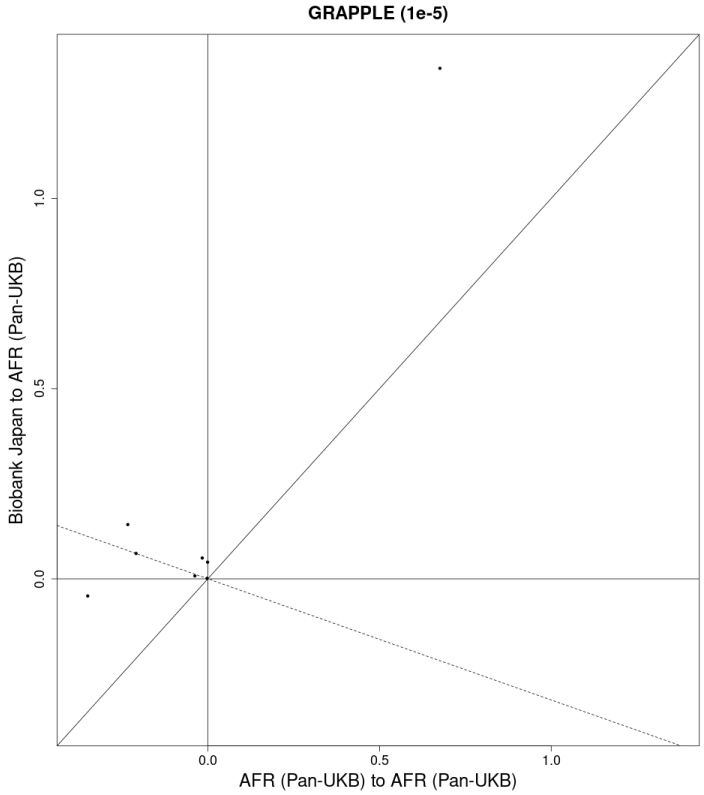

GRAPPLE (1e-5)
UK Biobank to Taiwan Biobank
Taiwan Biobank to Taiwan Biobank
GRAPPLE (5e-8)
UK Biobank to Taiwan Biobank
Taiwan Biobank to Taiwan Biobank
GRAPPLE (1e-5)
AFR (Pan-UKB) to UK Biobank
UK Biobank to UK Biobank
GRAPPLE (5e-8)
AFR (Pan-UKB) to UK Biobank
UK Biobank to UK Biobank


GRAPPLE (1e-5)
GRAPPLE (5e-8)
GRAPPLE (1e-5)
GRAPPLE (5e-8)
Biobank Japan to UK Biobank
FinnGen to UK Biobank
UK Biobank to UK Biobank


GRAPPLE (1e-5)
GRAPPLE (5e-8)
GRAPPLE (1e-5)
GRAPPLE (5e-8)
Majority Hispanic to UK Biobank
SAS (Pan-UKB) to UK Biobank
UK Biobank to UK Biobank


GRAPPLE (1e-5)
Taiwan Biobank to UK Biobank
UK Biobank to UK Biobank
-2
-1
0
1


GRAPPLE (5e-8)
Taiwan Biobank to UK Biobank
UK Biobank to UK Biobank
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5


GRAPPLE (1e-5)
Biobank Japan to AFR (Pan-UKB)
AFR (Pan-UKB) to AFR (Pan-UKB)
0.0
0.5
1.0

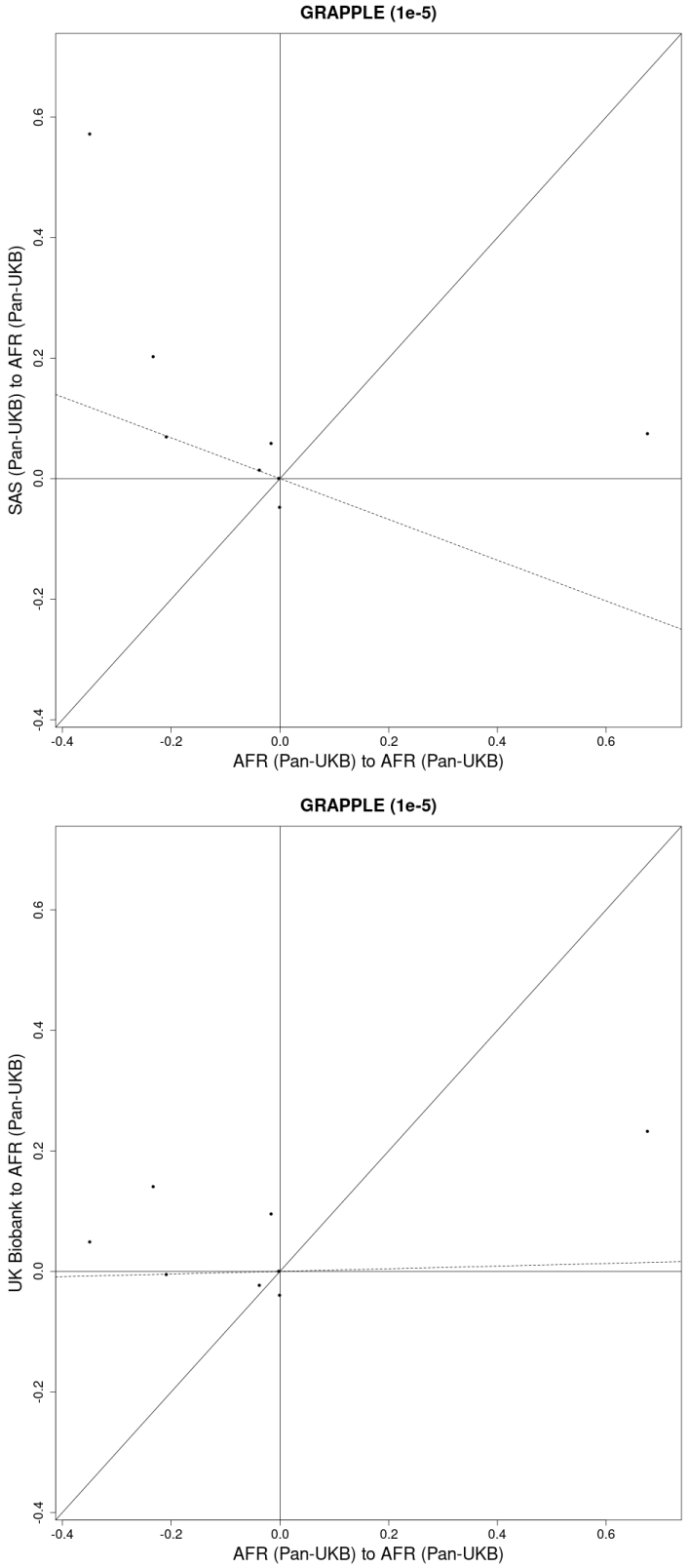


Figure S4: Cross-population vs. same-population 2SMR estimates for 23 population pairs at both instrument selection thresholds and 3 population pairs - using BBJ, South Asian (Pan-UKB), and UKB studies as exposures for African (Pan-UKB) outcomes - at only an instrument selection threshold of 1×10^{-5}. All estimates were calculated using GRAPPLE. The dashed line represents the SIMEX-estimated shrinkage coefficient for the population pair and instrument selection threshold.