

Manuscript Title: The power of genetic diversity in genome-wide association studies of lipids

Redactions – unpublished data

Parts of this Peer Review File have been redacted as indicated to maintain the confidentiality of unpublished data.

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referee #1 (Remarks to the Author):

The authors empirically show the improvement in analyses of blood lipid loci in genetically diverse individuals, over European studies. It is a well-written paper and an enjoyable read.

They show that the inclusion of multiple ancestries (20% non-European among 1.65 million individuals) leads to gains over European-only analyses in terms of fine-mapping resolution (decrease in number of putative causal SNPs). This agrees with many other studies that show improvements in the inclusion of diverse ancestries and contributes more refined results to the understanding of blood lipid levels. They also show the novel result that a polygenic score developed from trans-ancestry meta-analysis reduces the discrepancy previously observed with non-ancestry matched scores.

Data is composed of many ancestrally diverse cohorts. The authors clearly present their approach, including quality control of their data.

The abstract gives a clear overview of the results and both the introduction and discussion emphasise the importance of collecting data from diverse populations. The discussion gives nice tie-ins with existing efforts such as H3Africa and the All of Us program. The results are presented clearly and the supplementary results are comprehensive. I only have two comments for the authors to address below:

Comments:

1. Line 1082: The expression for the $\ln(\text{BF})$ basically says it is proportional to z^2 . This might be a typo. A BF compares the likelihood of the data under two different hypotheses and it is not clear how the form used does this. The authors should explain where this form comes from or correct the typo.
2. Line 1109: To make paper easier to read, it would be helpful to add a line to define PP3 and PP4 from coloc.

Referee #2 (Remarks to the Author):

The manuscript by Graham et al describes a genome-wide association study of blood lipids using an ancestrally diverse sample. The study identifies a large number of novel genome-wide significant associations with lipids. They also assess the performance of PRS and create credible sets.

This work will certainly become the new reference paper for lipid GWAS. The authors discover a novel associations with lipids and have the largest sample size to date. The summary statistics will be useful for other researchers. The novel genetic associations will provide the basis for follow up studies which may uncover the functional mechanisms.

Beyond that I am unable to see major new insights or innovations. The authors are stressing that their work demonstrates empirically the “power of genetically diverse individuals”. I would only

partially agree with that. They find that there is no advantage for discovery. They claim advantages for fine-mapping and polygenic risk scores. However, each of these analyses has weaknesses, the issues are described below, and therefore does not constitute the striking empirical evidence from a large-scale study demonstrating the advantages of diversity that one would have expected based on previous research.

I am sceptical about the robustness of the GWAS and fine-mapping findings for methodological reasons, outlined below.

It is disappointing that such a pro-diversity flagship paper from globally leading teams still fails in terms of inclusivity (individuals with admixed ancestry are excluded) and handles the highly important theme of diverse ancestry and ethnicity in a relatively undifferentiated way, both with respect to the science as well as the wider implications.

Finally, it is not clear how much of the data are actually novel.

Major comments:

The authors do not find that the discovery of novel genetic associations is improved by the more diverse sample. The discovery power is proportional to sample size in their study. That is a surprising finding given that differences in variant frequencies alone should increase power for a subset of the genetic variation. I wonder if this is due to the genotyping arrays that were used. If they are designed to optimise coverage of genetic variation in European ancestry samples, potential gains may diminish. Hence, the paper's message may be misleading.

Why were samples with admixed ancestry excluded in this work? Similar arguments for their inclusion apply as for non-European groups.

The paper handles the theme of diverse ancestry and ethnicity in a relatively undifferentiated way. For example, the authors do not give any consideration to the origin of their samples. Their study table in the supplement does not even list the country of recruitment. Looking at their study list, I get the impression that none (or almost none) of their "African" samples have been recruited from an African country. Most East Asian studies, on the other hand, appear to be from East Asian countries. Could this impact on the results? Do we actually know that their PRS results are equally applicable to people in a given country from Sub-Saharan Africa as they are to those UK Biobank participants? There is evidence that factors linked to country can impact on lipid associations (<https://doi.org/10.1038/s41467-019-12026-7>). What about the wider implications of this imbalance in terms of global health and translation of genomic findings?

Methodological issues

I am concerned about the validity of the fine-mapping results and credible sets in particular. This work is largely based on summary statistics. To create credible sets the authors needed LD information. Those were not derived from the studies themselves but from reference sets. Some of these reference sets are very different in terms of their population structure compared to the discovery GWAS. For example, the UK Biobank "White British" group is a relatively homogeneous group and excludes other ancestrally European samples. The discovery GWAS includes diverse ancestrally European groups and large numbers of individuals from USA which will be less homogeneous. Even slight LD mismatch can have a large impact on this type of summary statistics based fine mapping.

The GWAS was based on summary statistics. Variants were included if they had an imputation accuracy of 0.3. This is a very low value and includes variants with high levels of uncertainty.

Variants with such low levels of reliability should not be reported as novel findings especially since this study has not done any replication. It is not clear how reliable the reported hits are because the authors do not report the imputation accuracy for them.

Applying Hardy-Weinberg to genetic variants in African American and other genetically heterogeneous groups with some level of admixture is not a sensible application. The HW test is used to pick up technical errors. However, in those groups many variants will be out of Hardy Weinberg just due to differences in frequency between ancestral populations which is expected and no reason for exclusion.

The QC information is very limited. What QC was applied pre-imputation? Were there differences between studies? The authors also excluded entire cohorts. Firstly, it would be good to know which ones those were. Secondly, one of the reasons was that the effect estimate for established loci was not in the same direction as in the discovery study. That could be an issue because it would systematically disadvantage non-EUR studies which are likely to be less consistent with the EUR discovery ones and the authors could impact on some of their findings such as PRS performance which is likely to appear better when studies not confirming to their preconceptions have been removed.

The double GC correction is concerning. GC correction itself is usually considered unnecessarily conservative. GC correction twice seems unreasonably conservative. My concern is that the analyses did not sufficiently adjust for population stratification. Without this correction were there large numbers of likely false positive associations? This crude approach is not an ideal solution because some false positives might remain and the problem would also impact on PRS. It would be better to address the issue working directly with the source studies.

For the comparison of the performance of different PRS, the impact of ancestry and sample size are confounded. The best performing trans-ancestry PRS also has the largest discovery sample.

Data sharing and availability of results for future research is important. Previous GLGC releases have not provided ancestry specific results which could have been particularly useful for follow up research. It would have been desirable to make the downloads available to the reviewers so it can be evaluated if they fulfil the spirit of the open access policies of the journal.

Referee #3 (Remarks to the Author):

Graham et al conduct the largest GWAS yet of blood lipid levels with a focus on LDL-C, reaching an impressive sample size of ~1.65 million people. To amass these large numbers, the authors include participants with ancestries at a rate typical of the more general GWAS field (i.e. 80% of GWAS participants in this study and across the field are of European descent). As expected from a study this large, the authors identify a sizable number of novel associations from both ancestry-specific and trans-ancestry analyses. They also show that fine-mapping resolution is improved from multi-ancestry analyses and that polygenic score accuracy increases particularly for populations underrepresented in genetics. This study represents a significant amount of work and collaboration, and these results are likely of high interest particularly to the statistical genetics field. This large effort notwithstanding, the findings here are expected based on past studies (e.g. Klarin et al 2018 Nat Gen, Sakaue et al 2020 Nat Med, Martin et al 2019 Nat Gen, Mahajan et al 2018 Nat Gen, etc), with somewhat limited technological innovation and new biological insights into the phenotypes of interest here or related diseases. This work may therefore be a better fit for a field-specific journal such as Nature Genetics or similar.

Major comments:

1. Polygenic score analysis. Compared to a study from 10 years ago, this study has contributed to modestly improved genetic prediction accuracy for LDL in European ancestry groups (R^2 from 0.12 to 0.15) but also to much more improved accuracy for African Americans. A more sensible comparison to benchmark improvements in genetic prediction accuracy of LDL for African Americans, however, would be using GWAS summary statistics to compute scores from the more recent Klarin et al Nat Gen 2018 study. How does prediction accuracy from this study compare to these more recent studies, perhaps also including Willer et al 2013?

2. Figure 1B shows number of index variants as a metric for the relative contributions of ancestries by sample size, which is less informative than it could be as all primarily non-European ancestry populations are an order of magnitude smaller than Europeans, leading to effectively two clusters of points (the zoom in via Fig S7 is helpful). Perhaps a more helpful assessment would be to measure the effective sample size increase as populations are added based on GWAS power (perhaps with # significant loci or alternatively mean chi-square statistics—less impacted by LD differences, or some other metric of GWAS power). It would be helpful to understand how genomic discovery for these traits differs across populations at the ~50k and ~100k sample size. The discovery curve is described on lines 798 as linear, but a curve drawn between 2 clusters of points is a line and it's hard to know what the dynamics here are. Perhaps also plotting the subsets described in the polygenic score section of 100k, 200k, and 400k individuals would be helpful.

3. Biological interpretation. With increasing functional resolution becoming increasingly available, I was surprised not to see any downstream follow up, even if just cell type or other functional enrichment analyses for association signals.

4. The authors describe that polygenic scores are equally accurate in African American and European ancestry populations after a tuning strategy in African Americans that they describe as selecting SNPs in Europeans and effects in African Americans. This approach has been described previously and was recently shown to retain population stratification issues, for example seen previous in the GIANT study (Zaidi & Mathieson, bioRxiv following from Berg et al and Sohail et al, eLife 2019). Have the authors tested these biases, for example by looking at how effect sizes correlate with PCs in a global reference panel like 1000 Genomes and by looking across admixture fractions (if admixed samples weren't excluded per comment #1) to test trends in % variance explained with ancestry? Given the trans-ancestry nature of the study, it would be helpful to more broadly understand whether and the extent to which there is residual population stratification in GWAS summary statistics.

5. Phenotype. The authors describe adjusting the phenotype by statin use (reduce LDL-C levels proportionally by a factor 0.7). Is natural variation in the presence of medications consistent with a linear reduction, particularly across different races/ethnicities? Is dose dependence expected? Are associations consistent among those on versus not on statins? Some additional citations here would be helpful. Because statins appear to not be prescribed at the same level across ethnicities, its impact on trans-ancestry analyses are important.

6. I have concerns about some of the underlying data. The PCs in Supp Fig 1 have an error, as PC1 separates EUR from EAS rather than AFR from all other populations, inconsistent with population genetics. Perhaps a labeling issue? Additionally, it is unclear what is shown—have most admixed populations been removed altogether? The lack of a cline on the PCs is unexpected for admixed populations described, including African American and Hispanic/Latino populations. Following from these, I have concerns about Figure 1. While tSNE plots are purely intended for visualization and have no interpretable significance, their use with genetic data in ancestrally biased cohorts are misleading. This plot is no exception as it would most intuitively suggest that European ancestry populations have more variation than all other ancestries, which is clearly not

the case. PCA plots are preferable as these generate the covariates actually used in analyses and provides a more straightforward interpretation of the data, albeit with more dimensions. UMAP also performs dimensionality reduction with PCs but is less sensitive to hyperparameter tuning than tSNE, better preserving the global structure of PCs.

Minor comments:

1. Data availability. It is critical that ancestry-specific as well as trans-ancestry GWAS summary statistics be made available. A dummy link is included but does not yet work—there should be a process to ensure these are accessible. Relatedly, it can be helpful to the field to be able to test analyses by holding out a cohort such as the widely accessed UK Biobank data for testing methods, assumptions, etc as described in the polygenic score section. Will the authors make cohort-level statistics available or possibly a set of summary statistics withholding UK Biobank upon publication?
2. It would be helpful to have a more prominent summary of number of associated loci that are trait- and ancestry-specific versus shared.
3. Relative contributions of ancestries by sample size. Some of the explanations could be clearer. For example, in lines 779-782, the authors attribute the larger number of novel associations identified in AFR to the allele frequency differences between African and European populations. Another major contributor is the larger effective population size of the African populations (i.e. more genetic variants overall in these populations), as a population with high sample size that has been isolated could be equally differentiated but have fewer genetic variants overall.
4. I'm not sure how to compare the #'s of unique loci in line 779-882 with the population-specific variants that do not remain significant (lines 793-796 – shouldn't these be strictly less?).
5. In lines 805-808, it could be clearer that variant effect size *estimates* can be different because of LD patterns, but that in this scenario the causal effect itself can remain the same.
6. The authors describe a very large number of associated loci. What fraction of the genome do independent genome-wide significant associations cover?

Author Rebuttals to Initial Comments:

Referee #1 (Remarks to the Author):

The authors empirically show the improvement in analyses of blood lipid loci in genetically diverse individuals, over European studies. It is a well-written paper and an enjoyable read.

[Thank you for this comment, we spent substantial time polishing the message and readability.](#)

They show that the inclusion of multiple ancestries (20% non-European among 1.65 million individuals) leads to gains over European-only analyses in terms of fine-mapping resolution (decrease in number of putative causal SNPs). This agrees with many other studies that show improvements in the inclusion of diverse ancestries and contributes more refined results to the understanding of blood lipid levels. They also show the novel result that a polygenic score developed from trans-ancestry meta-analysis reduces the discrepancy previously observed with non-ancestry matched scores.

Thank you. We also believe that the lack of discrepancy in accuracy of the polygenic score in diverse ancestry groups is the major finding, and a relief for geneticists that future prevention strategies can avoid exacerbating racial inequities in disease.

Data is composed of many ancestrally diverse cohorts. The authors clearly present their approach, including quality control of their data.

Thank you. Unlike previous GWAS studies, we started with an extremely detailed analysis plan to contributing cohorts (including specific command lines and software to use for various stages), which resulted in a much cleaner dataset with many fewer QC issues than observed in previous meta-analyses we've worked on.

The abstract gives a clear overview of the results and both the introduction and discussion emphasise the importance of collecting data from diverse populations. The discussion gives nice tie-ins with existing efforts such as H3Africa and the All of Us program. The results are presented clearly and the supplementary results are comprehensive. I only have two comments for the authors to address below:

Thank you for noticing the detail in the Supplementary Tables. We hoped to provide a very clear presentation of all the data.

We address both of your helpful comments below.

Comments:

1. Line 1082: The expression for the $\ln(\text{BF})$ basically says it is proportional to z^2 . This might be a typo. A BF compares the likelihood of the data under two different hypotheses and it is not clear how the form used does this. The authors should explain where this form comes from or correct the typo.

Thank you for the clarifying question, we have revised the text to be more explanatory and we have added a citation to the text referencing the original paper. We include below a more detailed explanation of how the Bayes' factor may be approximated from the effect sizes and standard errors. Please see new text excerpted from the manuscript:

For a test statistic, X , that has an approximate chi-square distribution with K degrees of freedom under the null hypothesis, an approximation to the Bayes' factor in favour of the alternative hypothesis was derived by Schwarz (1978) and is given by

$$\text{BF} = \exp \left[\frac{X - K \ln N}{2} \right],$$

where N is the sample size. In this application, $X = Z^2$ and $K = 1$, such that

$$\text{BF} \propto \exp \left[\frac{Z^2}{2} \right].$$

New reference: Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, 6, 461-464.

Bayes factors^{51,52} (BF) for each variant in the ancestry-specific meta-analysis were approximated by:

$$\ln(BF) \propto 0.5 \frac{\beta^2}{SE^2}$$

where β and SE are the effect sizes and standard errors from the RAREMETAL meta-analysis.

2. Line 1109: To make paper easier to read, it would be helpful to add a line to define PP3 and PP4 from coloc.

Good point. We have added a definition of PP3 and PP4 to the colocalization methods.

The eQTL and GWAS signals (based on p-values from MR-MEGA) were considered to be colocalized if $PP3 + PP4 \geq 0.8$ and if $PP4/(PP3+PP4) > 0.9$, where PP3 is the probability of two independent causal variants while PP4 is the probability of a single, shared causal variant.

Referee #2 (Remarks to the Author):

The manuscript by Graham et al describes a genome-wide association study of blood lipids using an ancestrally diverse sample. The study identifies a large number of novel genome-wide significant associations with lipids. They also assess the performance of PRS and create credible sets.

This work will certainly become the new reference paper for lipid GWAS. The authors discover a novel associations with lipids and have the largest sample size to date. The summary statistics will be useful for other researchers. The novel genetic associations will provide the basis for follow up studies which may uncover the functional mechanisms.

Thank you. We agree that there is much to be discovered in this dataset and future functional studies will help explore the depth of the new findings here.

Beyond that I am unable to see major new insights or innovations. The authors are stressing that their work demonstrates empirically the “power of genetically diverse individuals”. I would only partially agree with that. They find that there is no advantage for discovery. They claim advantages for fine-mapping and polygenic risk scores. However, each of these analyses has weaknesses, the issues are described below, and therefore does not constitute the striking empirical evidence from a large-scale study demonstrating the advantages of diversity that one would have expected based on previous research.

We note the reviewer's concerns about the analysis aspects and rebut each one below. There is power in diverse samples in fine-mapping and PRS portability.

I am sceptical about the robustness of the GWAS and fine-mapping findings for methodological reasons, outlined below.

We hope we have alleviated the outlined concerns with revisions to the text and more descriptions of the results for each point below. Our manuscript team has substantial depth and expertise in genetic discovery, statistical genetics analysis and trans-ancestry analyses which is reflected in the first reviewer's comment 'The authors clearly present their approach, including quality control of their data.'

It is disappointing that such a pro-diversity flagship paper from globally leading teams still fails in terms of inclusivity (individuals with admixed ancestry are excluded) and handles the highly important theme of diverse ancestry and ethnicity in a relatively undifferentiated way, both with respect to the science as well as the wider implications.

Thank you for acknowledging GLGC as a globally leading team. We agree that improved genetic diversity in association studies is key to future utility of discovery efforts, but also wanted to be thorough in enumerating and evaluating the actual gain among 3 different types of analysis. To refine the message that diversity of study populations is important, we now enumerate the proportion of significant loci that would not have sufficient power to be detected in other ancestry groups (new Figure 1). We find that African-ancestry identified index variants are most frequently ancestry-specific compared to other ancestry groups, with 39% of African-ancestry index variants having <80% power to be detected in an equivalent number of individuals from any other included ancestry. We felt it was important to quantify, in a large sample size, the gains from diverse samples in terms of these three metrics: loci discovered (and now adding independent variants discovered), fine-mapping and PRS portability and utility. We believe that these are important findings for future participant recruitment and study design.

Finally, it is not clear how much of the data are actually novel.

Thank you for noting this lack of clarity. As noted by reviewer 3, the most recent large-scale lipid GWAS from MVP (Klarin et al, 2018) had a sample size ~6x smaller, discovered 6x fewer novel loci, and did not examine PRS portability at all. Furthermore, the number of non-European samples studied here is 4x larger (330k vs 81k) relative to the Klarin et al. study, and we here additionally assess PRS portability in an additional 296k diverse individuals, allowing for a much better assessment of ancestry-specific findings. We also generated a 'square' dataset where all populations completed genome-wide imputation prior to analysis, whereas previous GLGC studies published after 2010 (and the GLGC replication included in Klarin et al.) were a mix of targeted follow-up of some regions (Metabochip, Willer et al 2013) or coding variant arrays (exome chip, Liu et al 2017 and Lu et al 2017).

We have added statements to the text to clarify that the meta-analysis data reported here has not previously been published. Some of the studies may have been published individually (see Supp Table 1). However, relative to the most recent imputation-based meta-analysis from GLGC, the present meta-analysis represents a 35-fold increase in the number of variants through imputation with 1000 Genomes and/or the Haplotype Reference Consortium and a 16.5-fold increase in sample size (Teslovich et al Nature 2010, HapMap imputed). To state the novelty in the results, we now include in the Abstract:

"...237 novel loci from ancestry-specific analysis and an additional 120 novel loci from trans-ancestry analysis."

This is the first demonstration of a PRS that works as well in African and African-American cohorts as it does in European cohorts, for any trait. We also examine differences in performance within African continental population groups. Of note, racial/ethnic differences in PRS prediction remain a key barrier to clinical use; our mitigation of this important disparity provides a roadmap for other complex diseases representing an overall conceptual advance for PRS. We provide guidance to the field on how to achieve that for other traits and diseases by leveraging a very large-sample size and a well-powered polygenic quantitative trait. We demonstrate that both the variant selection and ancestry-specific effect sizes are critical to the PRS performance, although the latter more so. We also demonstrate that studies in well-powered African-American cohorts provide strongly portable PRS scores, even to West and South African cohorts with no European admixture.

This study has important implications for both cohort development and PRS methods development across a range of traits and diseases, suggesting that existing PRS tools are suitable for prediction in other ancestries so long as sufficient sample sizes representative of diverse population groups are reached. Lipid levels are an exemplar phenotype because of the heritable nature, polygenicity, power to detect association (especially relative to binary disease phenotypes), number of global studies with lipid levels collected, and important role in heart disease.

Major comments:

The authors do not find that the discovery of novel genetic associations is improved by the more diverse sample. The discovery power is proportional to sample size in their study.

That is a surprising finding given that differences in variant frequencies alone should increase power for a subset of the genetic variation.

Differences in allele frequencies between ancestries will sometimes increase power, and sometimes decrease power, for a given ancestry, depending on whether the minor allele frequency is higher or lower. We expect that the relative increase in number of new loci, given the number already discovered, may be slightly higher in understudied ancestries, but that is balanced by the very large sample sizes in European ancestries and the vast number of nearly-genome-wide-significant loci in these well powered populations.

However, all populations with sufficient sample size to detect association will have some ancestry-specific significant associations due to these differences. We have now included additional analyses to better address the question of genetic discovery between ancestry groups and how this varies with increasing sample size. We now show in **Figure 1** and **Supplementary Figure 3** the differences in effect size and allele frequency for variants identified between ancestry groups. We also show in **Figure 1F** the proportion of index variants from each ancestry that would likely be detected in other ancestry groups (with estimated power > 80%). As noted by the reviewer, we find that African-ancestry index variants are less likely to be detected in an equivalent sample size of non-African ancestry individuals due to power differences. In **Supplementary Figure 7**, we show the trends in locus discovery with varying European sample size (including subsets of approximately 100K, 200K, 400K, 600K, 800K, or 1.2 million individuals). We include below the edited text from the manuscript, which we believe now better reflects the nuances in GWAS discovery based on currently available sample sizes and ancestral groups.

*Next, we compared the number of loci identified between ancestry groups per 100,000 participants (**Figure 1**). The European and trans-ancestry analyses identified the greatest absolute number of total and novel loci, reflecting their large sample size. However, analyses of non-European ancestry groups identified a greater number of associations relative to the number of included individuals and African-ancestry analyses identified the third highest proportion of novel loci, reflecting the value of including diverse groups even if available sample sizes are smaller. We then estimated the proportion of index variants within each ancestry group that would likely be identified from analyses of an equivalent sample size in other ancestry groups (**Figure 1**). Of the index variants that were significant in the African-ancestry meta-analysis, we estimate that 39% would not reach significance in a meta-analysis of an equivalent number of individuals in any other ancestry group, due to an estimated power of <80%. To further examine trends in the number of loci with increasing study sample size, we compared both the number of identified loci and the mean chi-squared value relative to meta-analysis sample size across ancestry groups (**Supplementary Table 7**, **Supplementary Figure 7**). Both the number of identified loci and the mean chi-squared value were generally linearly*

*related to sample size as ancestry groups were meta-analyzed. Within the European ancestry group, however, the incremental increase in either the number of loci or chi-squared value per increase in sample size was attenuated at the largest sample sizes (**Supplementary Figure 7**). This suggests that once sufficiently well-powered sample sizes are reached within a given ancestry group, increasing sample sizes of other previously under-represented groups will most advance GWAS discovery.*

We also think it's an important question – just how many ancestry-specific variants can we expect to find in understudied ancestries? We attempted to quantify this with new Figure 1.

I wonder if this is due to the genotyping arrays that were used. If they are designed to optimise coverage of genetic variation in European ancestry samples, potential gains may diminish. Hence, the paper's message may be misleading.

We agree that past array design may be biased towards European variant coverage. Future meta-analyses incorporating more diverse reference panels (including TOPMed) and improved array designs will be essential to identify the full spectrum of variation across diverse individuals globally. However, many of the participating studies in the present analysis have used semi-custom arrays that are optimized for non-European populations. For example, the Million Veteran Program array (the largest contributing cohort for African American and Hispanic individuals) included an additional 50,000 markers designed to improve coverage and imputation quality of variants from African-ancestry populations [Hunter-Zinck et al. PMID:32243820] and the China Kadoorie Biobank used a custom array designed to improve coverage in Chinese individuals. Therefore, while we agree that better array designs and imputation panels will further improve future discovery efforts, we do not believe that the findings of this study are an artifact of European-centric arrays. Genotyping arrays and imputation panels for participating cohorts are included in **Supplementary Table 1**. We have added the following sentence to the text in recognition of this:

"Further gains in the depth and number of sequenced individuals of diverse ancestries available for array design and imputation into diverse cohorts may additionally improve discovery of novel variants and loci in diverse cohorts, particularly variants absent from other ancestries."

Why were samples with admixed ancestry excluded in this work? Similar arguments for their inclusion apply as for non-European groups.

We apologize for any confusion. We did not exclude admixed individuals – indeed individuals with cross-continental admixture (such as African-American or Black British) constitute ~5% of our study participants. We have gone carefully through the text to see how this misunderstanding may have occurred. We have clarified the allele frequency plot (**Supplementary Figure 1**) text to state that the tSNE plot of cohort-level data is from summary level information and therefore not expected to mirror standard individual-level data principal component plots, as we believe this may have been the source of confusion. We have also edited the main text to specify the origin of African-ancestry cohorts:

*In the present study, we performed a meta-analysis of 162 GWAS for 5 lipid levels in up to 1.65 million individuals from 5 ancestral-groups from around the world (AdmAFR; Admixed African or African individuals recruited from the US, UK, and Africa, HIS: Hispanic, EUR: European, EAS: East Asian and SAS: South Asian, **Supplementary Table 1**).*

The paper handles the theme of diverse ancestry and ethnicity in a relatively undifferentiated way. For example, the authors do not give any consideration to the origin of their samples. Their study table in the supplement does not even list the country of recruitment. Looking at

their study list, I get the impression that none (or almost none) of their “African” samples have been recruited from an African country. Most East Asian studies, on the other hand, appear to be from East Asian countries. Could this impact on the results?

We thank the reviewer for the suggestion to include country of origin information. We were primarily focused on ancestry as opposed to country of residence since we are primarily interested in studying genetic factors related to blood lipid levels.

Country of recruitment has now been collected from all 280 cohorts and added to **Supplementary Table 1**. We strongly agree that individuals who represent diverse populations worldwide need to be included in GWAS to ensure that results from these studies be equitably used in the future, and have added discussion of the importance of such initiatives (e.g. H3Africa and All of Us) in the manuscript:

“In this respect, the results of this study strongly support recent efforts such as the All of Us program and H3Africa to extensively diversify genetic studies and provides compelling evidence that these efforts will lead to accelerated biological and pharmaceutical translation of findings from genetic discovery studies, and possibly a reduction in existing health disparities through the application of genetic risk prediction algorithms to clinical practice that perform equally well across all major race/ethnic groups.”

The African-ancestry cohorts in the discovery GWAS include African-American individuals from the US, African-ancestry individuals in the UK biobank, and one cohort from South Africa. It's substantially more difficult and complicated to collect large cohorts from Africa with informed consent, DNA, genotyping and lipid phenotypes, thus the relative informativity of PRS in African cohorts from GWAS of African-admixed individuals is highly relevant to future studies. Although differences in diet, frequency of obesity and other environmental risk factors will be different by country of origin and may impact which associations are observed, it is important to consider both genetic ancestry and country (or continent) of residence.

The cohorts are included in **Supplementary Figure 1** and **Supplementary Table 1** and are now explicitly listed in the main text to avoid any possible confusion. We also more clearly delineate individuals of African-admixed ancestry and African individuals throughout the text.

We also thank the reviewer for suggesting we analyze the prediction of the lipid PRS for more specific ancestry groups, such as studies recruited from African populations.

Do we actually know that their PRS results are equally applicable to people in a given country from Sub-Saharan Africa as they are to those UK Biobank participants? There is evidence that factors linked to country can impact on lipid associations (<https://doi.org/10.1038/s41467-019-12026-7>). What about the wider implications of this imbalance in terms of global health and translation of genomic findings?

We now include testing of the PRS in 295,577 individuals, including cohorts from South, West and East Africa, from two different studies. With the addition of two cohorts recruited from Africa, we have now added the following text:

“We then evaluated the optimized PRS scores in 8 cohorts of individuals (N=295,577, Supplementary Table 14) not included in the discovery GWAS, from 6 ancestral groups: Continental Africa (2,452 East Africa, 4,972 South Africa, 7,309 West Africa), African American (21,730), Hispanic American (7,669), East Asian (146,477), South Asian (15,242), Asian American (4,155), and European American (85,571).”

We show that the trans ancestry-derived PRS predicts LDL similarly well in African Americans, West Africans and those with solely European ancestry. We also include a discussion of the need for increased representation of other African-ancestry populations and a better understanding of potential population-specific and environmental differences between groups. We have also expanded our analysis to include testing the PRS in 5 ancestral groups to examine the correlation between PRS and PCs and have added an analysis of PRS prediction by proportion of African ancestry for the African-American individuals in MGI and MVP. We have included the updated **Figure 3** and relevant text below.

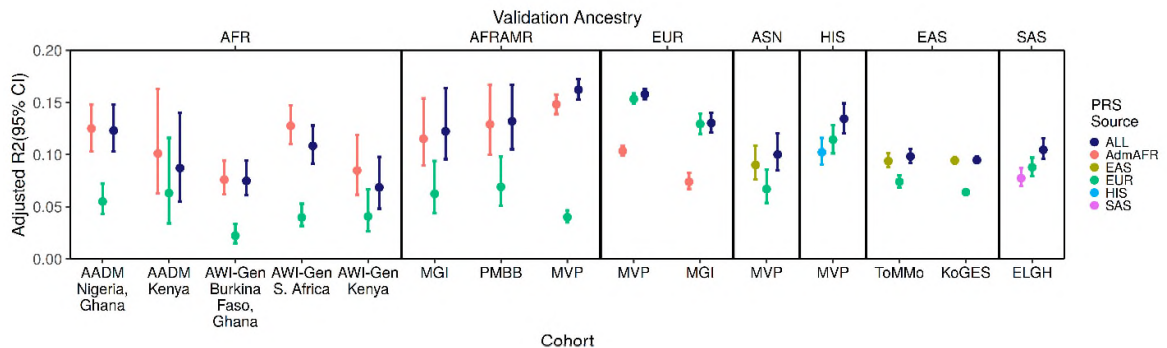


Figure 3: Trans-ancestry LDL-C PRS show similar performance across ancestry groups

Polygenic scores generated from trans-ancestry meta-analysis show equivalent or better performance across most ancestry groups relative to ancestry-specific PRS within each cohort, whereas European-specific scores show less transferability. Adjusted R^2 is calculated with the risk score as a predictor of LDL-C in a linear model with covariates. AFR: African, AFRAMR: African American, ASN: Asian American

New text from results section:

*“We then evaluated the optimized PRS scores in 8 cohorts of individuals (N=295,577, **Supplementary Table 14**) not included in the discovery GWAS, from 6 ancestral groups: Continental Africa (2,452 East Africa, 4,972 South Africa, 7,309 West Africa), African American (21,730), Hispanic American (7,669), East Asian (146,477), South Asian (15,242), Asian American (4,155), and European American (85,571).*

*The score developed from trans-ancestry meta-analysis consistently showed the best or near-best performance across ancestry groups (**Figure 3, Supplementary Table 15**), with improved or equivalent prediction relative to ancestry-matched scores. This was especially evident when limited sample sizes of ancestry-matched cohorts were available in the GWAS, as was the case for HIS and SAS. In contrast, prediction in South and East African individuals was slightly lower (by 0.01-0.02, **Supplementary Table 15**) for the trans-ancestry score relative to the African-ancestry derived score within each of the AADM and AWI-Gen regional subgroups. We further assessed the prediction of the scores relative to the proportion of African admixture in MGI and MVP participants. PRS prediction was similar for participants across all quartiles of African ancestry proportions (**Supplementary Figure 16**). Previous studies have suggested that population stratification may influence the predictive ability of polygenic scores across diverse populations²⁹. We tested for correlation between the trans-ancestry polygenic score and principal components (PCs) in 1000 Genomes individuals. Significant correlation was observed between the trans-ancestry polygenic score and principal components 1 and 3 only (p -value < 0.0025; 0.05/20 tested PCs), indicating that the minimum of 4*

PCs used in our polygenic score analysis were sufficient to correct for this (Supplementary Table 16).“

New text from discussion section:

“However, the LDL-C polygenic score showed greater variability in prediction of LDL-C for cohorts within Africa than it did in African Americans. Mean lipid levels within each cohort also exhibited greater variation between the continental African cohorts compared to all other ancestry groups (AFR range: 71.9-131.5 mg/dL, all others: 113.1-129.2 mg/dL). This suggests that sufficient representation of individuals and genetic variation from the risk prediction population improves performance, even for trans-ancestry derived scores. Therefore, additional studies are needed to better understand both the genetic and environmental factors influencing LDL levels within African populations and elsewhere. This underscores the need to ensure that future GWAS studies are representative of global populations. “

I am concerned about the validity of the fine-mapping results and credible sets in particular. This work is largely based on summary statistics. To create credible sets the authors needed LD information. Those were not derived from the studies themselves but from reference sets. Some of these reference sets are very different in terms of their population structure compared to the discovery GWAS. For example, the UK Biobank “White British” group is a relatively homogeneous group and excludes other ancestrally European samples. The discovery GWAS includes diverse ancestrally European groups and large numbers of individuals from USA which will be less homogeneous. Even slight LD mismatch can have a large impact on this type of summary statistics based fine mapping.

In designing the present study, we aimed to collect covariance matrices from each cohort to allow for exact conditional analysis using LD information from individual studies within the raremetal program used for meta-analysis, and indeed collected 260 terabytes of LD matrices from over half of the 280 cohorts. However, this was impossible to generate and export for many of the largest cohorts (including the Million Veteran Program) due to restrictions on data sharing and processing, which therefore necessitated the use of reference LD information.

We calculate the credible sets based on Bayes factors derived from the observed effect size and standard errors in the ancestry-specific or trans-ancestry analyses. To ensure consistent sample size across variants tested, we did not meta-analyze with any prior datasets (such as exome chip, metabochip follow-up or HapMap-based imputation panels). LD reference panels (using all ancestries from 1000 Genomes) in the context of generating credible sets were used only to identify shared association signals based on the linkage disequilibrium between index variants from different traits. We did this to avoid double-counting improvement in fine-mapping from overlapping association signals, which may otherwise overstate the estimated reduction in credible set size. LD information was not otherwise needed to generate credible sets. We have added an additional citation to the text referencing the formula used and include the derivation below:

For a test statistic, X , that has an approximate chi-square distribution with K degrees of freedom under the null hypothesis, an approximation to the Bayes' factor in favour of the alternative hypothesis was derived by Schwarz (1978) and is given by

$$BF = \exp \left[\frac{X - K \ln N}{2} \right],$$

where N is the sample size. In this application, $X = Z^2$ and $K = 1$, such that

$$BF \propto \exp\left[\frac{Z^2}{2}\right].$$

Schwarz, G. (1978). *Estimating the dimension of a model.* *Ann. Statist.*, 6, 461-464.

The GWAS was based on summary statistics. Variants were included if they had an imputation accuracy of 0.3. This is a very low value and includes variants with high levels of uncertainty. Variants with such low levels of reliability should not be reported as novel findings especially since this study has not done any replication. It is not clear how reliable the reported hits are because the authors do not report the imputation accuracy for them.

We appreciate this concern. The selected cutoff of 0.3 is a compromise between competing factors that reduce power: false negatives (excluding a truly-associated variant that is imperfectly estimated) and reduced effective sample size (for variants with imperfect imputation). Reduced imputation quality (i.e. noise in genotype dosages) reduces power but does not introduce false positives unless an unusual situation occurs where the variant imputation is of differing quality for those with high vs low measurements of the lipid phenotype.

We note that the cutoff, imputation $r^2 < 0.3$ for imputation quality, has been the standard used since the publication of several commonly used imputation tools (PMID: 22820512, 19543373) and is consistent with that used in many large, recent high-impact studies (e.g. Mahajan et al. *Nature Genetics* 2018 with $n \sim 900,000$ and Spracklen et al. *Nature* 2020 with $n \sim 430,000$). We have added a column to the supplementary tables with the imputation quality of each index variant in the UK Biobank, the largest contributing cohort. The median imputation score within the UK Biobank for lipid-associated variants from both ancestry-specific and trans-ancestry analysis was 0.995 (mean ancestry-specific: 0.966 and mean trans-ancestry: 0.969). The fraction of index variants with an imputation quality score > 0.8 in the UK biobank was 97%. The results also reflect that a large majority of our novel loci/index variants are very well-imputed. Furthermore, we did not include variants found in only a single study when identifying index variants. **Supplementary Tables 3 and 6** include the number of studies each index variant was observed in. We have added the following text to the manuscript:

“During the QC process, within each cohort we removed poorly imputed variants (info score or $r^2 < 0.3$), variants deviating from Hardy-Weinberg Equilibrium (HWE p -value $< 1 \times 10^{-8}$, except for MVP which used HWE p -value $< 1 \times 10^{-20}$), and variants with minor allele count < 3 . An imputation info score threshold of 0.3 was selected in order to balance the inclusion of variants across diverse studies while removing poorly imputed variants.”

Applying Hardy-Weinberg to genetic variants in African American and other genetically heterogeneous groups with some level of admixture is not a sensible application. The HW test is used to pick up technical errors. However, in those groups many variants will be out of Hardy Weinberg just due to differences in frequency between ancestral populations which is expected and no reason for exclusion.

Thank you for this comment. We chose to exclude poor quality variants based on deviation from Hardy-Weinberg equilibrium within each study to avoid false positive associations due to incorrect genotyping. We compared the number of variants removed by Hardy-Weinberg filtering within studies across ancestries and found that similar numbers of variants were

removed. For example, in the Multi-Ethnic Study of Atherosclerosis (MESA) cohort, 0.1% of variants in African Americans were removed by the Hardy-Weinberg filtering, compared to 0.4% in East Asians, 0.1% in Europeans, and 0.3% in Hispanic individuals. Within the Million Veteran Program, the largest contributing cohort for African American and Hispanic individuals, variants were filtered at a HWE p-value of 1×10^{-20} and we have updated this key point in the Methods text. The HWE p-value applied to other contributing studies was $p < 1 \times 10^{-8}$.

The QC information is very limited. What QC was applied pre-imputation? Were there differences between studies? The authors also excluded entire cohorts. Firstly, it would be good to know which ones those were. Secondly, one of the reasons was that the effect estimate for established loci was not in the same direction as in the discovery study. That could be an issue because it would systematically disadvantage non-EUR studies which are likely to be less consistent with the EUR discovery ones and the authors could impact on some of their findings such as PRS performance which is likely to appear better when studies not confirming to their preconceptions have been removed.

We agree with the reviewer that this could have been an issue if cohorts were excluded for failure to reproduce known loci, however, we did not do this. We have re-read the manuscript and revised any text that may have implied this. The 3 cohorts we excluded following QC were for issues related to improper imputation and trait transformation during analysis (Korea National Diabetes Program, EAS, N=1,468; Norfolk Island, EAS, N=276; Cebu Longitudinal Health and Nutrition Survey, EAS, N=1,782).

We have expanded the methods section to include the pre-imputation QC steps (cohort-specific information is included in **Supplementary Table 1**). Checking concordance of directions of effect for previously known loci revealed some important analysis issues amongst the 280 cohorts x 5 phenotypes, including switching of LDL and HDL in one cohort and identifying a cohort which had incorrectly aggregated multiple traits into a single lipid trait. These analysis issues were subsequently corrected by the cohort investigators and no cohorts were excluded from the final analysis for their inability to reproduce known loci. The updated text is included below:

Cohort level analysis

*Each cohort contributed summary statistics, imputation quality statistics, and analysis metrics for QC, following a detailed analysis plan (**Supplementary File 1**). Briefly, we requested that each cohort perform imputation to 1000 Genomes Phase 3 (1KGP3), with European cohorts additionally imputing with the Haplotype Reference Consortium (HRC) panel using the Michigan Imputation Server (<https://imputationserver.sph.umich.edu/index.html#!>) which uses Minimac software³⁷. Detailed pre-imputation QC guidelines were provided, these included removing samples with call rate < 95%, samples with heterozygosity > median + 3(interquartile range), PCA outliers within each ancestry group, and variants deviating from Hardy-Weinberg equilibrium (p -value < 10^{-6}) or with variant call rate < 98%. Analysis was to be carried out separately in each ancestry group and stratified by cases and controls where appropriate (i.e., based on ascertainment). Residuals were generated separately in males and females adjusting for age, age², principal components of ancestry, and any necessary study-specific covariates. Triglyceride levels were natural-log transformed before generating residuals. Inverse normalization was then done on the residual values. Individuals on cholesterol lowering medication had their pre-medication levels³⁸ approximated by dividing the LDL-C value by 0.7 and the TC value by 0.8. Association analysis of the residuals for the majority of cohorts was carried out using a linear mixed-model approach in *r*tests or with other similar software including BOLT-LMM³⁹, SAIGE⁴⁰, or deCode association software.*

Quality Control

Each input file was assessed for quality control using the EasyQC software⁴¹. We generated QQ plots by frequency bins, assessed trends in standard errors relative to sample size for each cohort, and checked allele frequency of submitted variants relative to their expected value based on the imputation reference panel. In addition, we checked that each cohort reproduced the expected direction of effect at the majority of known loci relative to the cohort sample size. Cohorts identified to have issues with the submitted files were contacted and corrected files were submitted or the cohort was excluded from meta-analysis. In some cases, cohorts had generated residuals of males and females together, without accounting for sex. In these cases, sex-stratified files were used for meta-analysis. During the QC process, within each cohort we removed poorly imputed variants (info score or $r^2 < 0.3$), variants deviating from Hardy-Weinberg Equilibrium (HWE p -value $< 1 \times 10^{-8}$), and variants with minor allele count < 3 . Summary statistics were then genomic-control (GC) corrected using the λ_{GC} value calculated from the median p -value of variants with MAF $> 0.5\%$. To capture all potential variants, summary statistics from cohorts that had submitted both HRC and 1KGP3 imputed files were joined, favoring variants imputed from HRC. For variants imputed by both panels, we observed that variants imputed from the HRC panel resulted in a higher imputation info score for 94% of variants when compared to the imputation info score from 1KGP3.

The double GC correction is concerning. GC correction itself is usually considered unnecessarily conservative. GC correction twice seems unreasonably conservative. My concern is that the analyses did not sufficiently adjust for population stratification. Without this correction were there large numbers of likely false positive associations? This crude approach is not an ideal solution because some false positives might remain and the problem would also impact on PRS. It would be better to address the issue working directly with the source studies.

We note that double GC-correction is fairly typical in the field for quantitative trait GWAS (e.g. MAGIC consortium, GIANT consortium, ICBP, etc.). We did not see evidence of uncorrected population stratification during QC of cohort level data, which included the use of linear mixed models and principal components when summary statistics were generated to correct for this. However, given the very large number of associated loci obtained with sample sizes ~ 1.6 million, we chose to be conservative in reporting associations by performing double-GC correction.

The distribution of the test statistics implies no significant systemic confounding. For example, within the trans-ancestry analysis, λ_{GC} values ranged from 1.09-1.14, which is within the expected range. All λ_{GC} values across traits and analyses are reported in **Supplementary Tables 2 and 5**. Summary statistics will be available upon publication and undoing the second GC correction would be trivial for anyone interested in that approach. Furthermore, because the GC-correction factor was relatively low for these phenotypes, the application or non-application of this correction would have very little substantive impact on the results, and likely no impact on the PRS.

We now emphasize the relatively low GC-correction factors in the methods:

Following meta-analysis, we identified loci based on a genome-wide significance threshold of 5×10^{-8} after GC correction using the λ_{GC} value calculated from the median p -value of variants with MAF $> 0.5\%$. The choice of double-GC correction was made to be most conservative and to minimize potential false-positive findings.

*Observed λ_{GC} values were within the expected range for similarly sized studies and are included in **Supplementary Tables 2 and 5**.*

For the comparison of the performance of different PRS, the impact of ancestry and sample size are confounded. The best performing trans-ancestry PRS also has the largest discovery sample.

We thank the reviewer for this important point. We agree that the issue of sample size could confound the comparison of African, European, and trans ancestry-derived risk scores. Therefore, we have now carried out an analysis of European sample sizes based on ~100K, 200K, and 400K individuals and a meta-analysis of 100K Europeans with the African-ancestry or African, East Asian, Hispanic, and South Asian ancestry groups to compare the role of sample size in the adjusted R^2 for prediction in both African American and European individuals (**Supplementary Figure 17**). We find that while increasing sample size of the European meta-analysis improves risk prediction in Europeans, this has a minimal impact on prediction in African Americans and that prediction in African Americans is most improved by the addition of non-European ancestry groups. The updated portion of the main text is included below.

*We then examined the role of GWAS sample size using the 2010 Global Lipids Genetics Consortium (EUR N=95,454, imputed with HapMap) LDL-C meta-analysis results⁵ alone or meta-analyzed with the present AdmAFR or non-European ancestry results. In addition, we generated subsets of the present European ancestry meta-analysis to yield 100K, 200K, or 400K individuals. For European ancestry individuals, we observed a 32% increase in the predictive power of LDL-C polygenic scores (as measured by adjusted R^2) with the nine-fold increase in sample size between the 2010 and present European-specific PRS (adjusted $R^2 = 0.12$ and 0.15 , respectively). However, the prediction of LDL-C in African-American individuals was relatively weak from an entirely European ancestry GWAS, irrespective of the sample size (adjusted $R^2 = 0.04$, **Supplementary Figure 17, Supplementary Table 15**).*

Data sharing and availability of results for future research is important. Previous GLGC releases have not provided ancestry specific results which could have been particularly useful for follow-up research. It would have been desirable to make the downloads available to the reviewers so it can be evaluated if they fulfil the spirit of the open access policies of the journal.

Thank you for your interest in our summary statistics and results from this study.

We strongly agree with the reviewer that performing ancestry-specific analyses and sharing the results is critically important to the field. Since the first GLGC publication in 2008, GLGC has publicly shared GWAS summary statistics from all publications by one-click download from our website [<http://lipidgenetics.org/#data-downloads-title>] including ancestry-specific results (for those studies with data from multiple ancestries: Lu et al Nat Genet 2017 and Liu et al Nat Genet 2017). For the results from this manuscript, we will share trans-ancestry and ancestry-specific meta-analysis results and PRS weights -- with and without UK biobank. The Million Veteran Program will also make all MVP cohort specific summary statistics available through dbGAP (to add to the racial/ethnic specific summary statistics already shared through dbGAP after the Klarin et al. publication in 2018). We expect this shared information to be useful for both analysis and methods development purposes. We will continue to describe and interpret the results from this work with a more detailed focus on characterizing candidate genes at each locus using a variety of different approaches, interpreting new biology, new loci, new pathways and new sex-specific findings from this work in a separate manuscript.

Referee #3 (Remarks to the Author):

Graham et al conduct the largest GWAS yet of blood lipid levels with a focus on LDL-C, reaching an impressive sample size of ~1.65 million people. To amass these large numbers, the authors include participants with ancestries at a rate typical of the more general GWAS field (i.e. 80% of GWAS participants in this study and across the field are of European descent).

To maximize information in the context of both diversity and statistical power, we sought to amass the largest absolute sample size of non-European (N~350k) and European (N~1.3m) samples possible. We downsampled the European cohort to evaluate the role of sample size in discovery of associated loci and variants and in the portability of PRS to diverse ancestries, and newly included two new cohorts from Africa.

As expected from a study this large, the authors identify a sizable number of novel associations from both ancestry-specific and trans-ancestry analyses. They also show that fine-mapping resolution is improved from multi-ancestry analyses and that polygenic score accuracy increases particularly for populations underrepresented in genetics. This study represents a significant amount of work and collaboration, and these results are likely of high interest particularly to the statistical genetics field. This large effort notwithstanding, the findings here are expected based on past studies (e.g. Klarin et al 2018 Nat Gen, Sakaue et al 2020 Nat Med, Martin et al 2019 Nat Gen, Mahajan et al 2018 Nat Gen, etc), with somewhat limited technological innovation and new biological insights into the phenotypes of interest here or related diseases. This work may therefore be a better fit for a field-specific journal such as Nature Genetics or similar.

Along with the multiple strengths of this study that the reviewer generously noted, this is the first demonstration that a multi-ancestry PRS performs equally well in continental Africans, admixed Africans *and* Europeans which is why we thought this work would be of interest to a broad audience. **It is an important finding that PRS can and should be equally informative in diverse populations, and that improving the diversity of the discovery GWAS is sufficient to achieve that so as not to exacerbate racial health inequities in future approaches aimed to prevent disease using genetics to predict at-risk individuals.**

Major comments:

1. Polygenic score analysis. Compared to a study from 10 years ago, this study has contributed to modestly improved genetic prediction accuracy for LDL in European ancestry groups (R^2 from 0.12 to 0.15) but also to much more improved accuracy for African Americans. A more sensible comparison to benchmark improvements in genetic prediction accuracy of LDL for African Americans, however, would be using GWAS summary statistics to compute scores from the more recent Klarin et al Nat Gen 2018 study. How does prediction accuracy from this study compare to these more recent studies, perhaps also including Willer et al 2013?

We thank the reviewer for the suggestion to compare the present risk score results with those generated from previous data sets. The Klarin et al. MVP manuscript from 2018 included ~57k African Americans but analyzed risk scores in European individuals using

weights from previously published GLGC summary statistics genotyped with an exome array. We have newly added an assessment of downsampled GWAS results from European samples of 100K, 200K, or 400K individuals.

We find that while increasing sample size of the European meta-analysis improves risk prediction in Europeans, this has a minimal impact on prediction in African Americans and that prediction in African Americans is most improved by the addition of non-European ancestry groups. The updated portion of the main text is included below.

*We then examined the role of GWAS sample size using the 2010 Global Lipids Genetics Consortium (EUR N=95,454, imputed with HapMap) LDL-C meta-analysis results⁵ alone or meta-analyzed with the present AdmAFR or non-European ancestry results. In addition, we generated subsets of the present European ancestry meta-analysis to yield 100K, 200K, or 400K individuals. For European ancestry individuals, we observed a 32% increase in the predictive power of LDL-C polygenic scores (as measured by adjusted R^2) with the nine-fold increase in sample size between the 2010 and present European-specific PRS (adjusted $R^2 = 0.12$ and 0.15 , respectively). However, the prediction of LDL-C in African-American individuals was relatively weak from an entirely European ancestry GWAS, irrespective of the sample size (adjusted $R^2 = 0.04$, **Supplementary Figure 17**, **Supplementary Table 15**).*

We compared the prediction of risk scores generated using the 2010 Teslovich et al. lipid GWAS association results, or downsampled European results from this study, as these included imputed variants genome-wide rather than the 2013 Willer et al paper, which used a two-stage design in which some variants were from the 2010 GWAS and some had a substantially larger sample size if genotyped on the metabochip array. We have included an analysis of polygenic scores calculated from the 2010 Teslovich et al results meta-analyzed with either African-ancestry or all non-European ancestries. We show that inclusion of larger numbers of European individuals improves prediction in Europeans but has minimal improvement on prediction in African Americans. We have included these results in **Supplementary Figure 17**.

2. Figure 1B shows number of index variants as a metric for the relative contributions of ancestries by sample size, which is less informative than it could be as all primarily non-European ancestry populations are an order of magnitude smaller than Europeans, leading to effectively two clusters of points (the zoom in via Fig S7 is helpful). Perhaps a more helpful assessment would be to measure the effective sample size increase as populations are added based on GWAS power (perhaps with # significant loci or alternatively mean chi-square statistics—less impacted by LD differences, or some other metric of GWAS power). It would be helpful to understand how genomic discovery for these traits differs across populations at the ~50k and ~100k sample size. The discovery curve is described on lines 798 as linear, but a curve drawn between 2 clusters of points is a line and it's hard to know what the dynamics here are. Perhaps also plotting the subsets described in the polygenic score section of 100k, 200k, and 400k individuals would be helpful.

We thank the reviewer for this good suggestion and we newly include meta-analyses with European subsets of approximately 100K, 200K, 400K, 600K, and 800K individuals and the 2010 Teslovich et al of 97k LDL meta-analysis results to further examine varying European sample sizes. Plots of chi-squared values compared to sample size for both common and low-frequency variants are shown in **Supplementary Figure 7C-D** for multi-ancestry analyses and in **Supplementary Figure 7E** within Europeans alone. We find that trends in

mean chi-squared value compared to mean sample size for each of the trans-ancestry meta-analyses are fairly linear, regardless of the ancestry groups included. However, we find that as the European sample size is increased there is a slight downward trend in both the mean chi-squared value per mean sample size and in the number of identified loci per sample size. We have updated Figure 1 to show the differing proportions of novel loci by ancestry group at currently available sample sizes (ie. much larger numbers of non-European individuals than European) and the relationship with novel and ancestry-specific GWAS discovery. We include the updated text below.

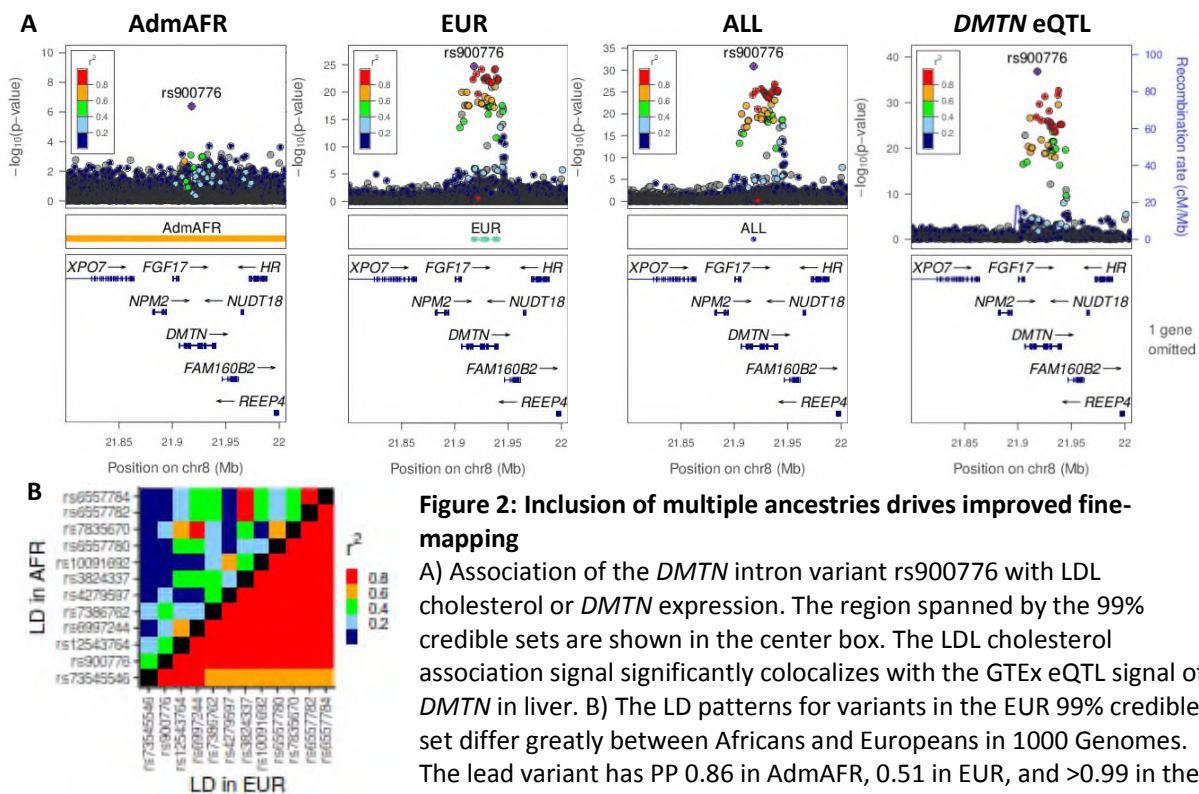
*Next, we compared the number of loci identified between ancestry groups per 100,000 participants (**Figure 1**). The European and trans-ancestry analyses identified the greatest absolute number of total and novel loci, reflecting their large sample size. However, analyses of non-European ancestry groups identified a greater number of associations relative to the number of included individuals and African-ancestry analyses identified the third highest proportion of novel loci, reflecting the value of including diverse groups even if available sample sizes are smaller. We then estimated the proportion of index variants within each ancestry group that would likely be identified from analyses of an equivalent sample size in other ancestry groups (**Figure 1**). Of the index variants that were significant in the African-ancestry meta-analysis, we estimate that 39% would not reach significance in a meta-analysis of an equivalent number of individuals in any other ancestry group, due to an estimated power of <80%. To further examine trends in the number of loci with increasing study sample size, we compared both the number of identified loci and the mean chi-squared value relative to meta-analysis sample size across ancestry groups (**Supplementary Table 7, Supplementary Figure 7**). Both the number of identified loci and the mean chi-squared value were generally linearly related to sample size as ancestry groups were meta-analyzed. Within the European ancestry group, however, the incremental increase in either the number of loci or chi-squared value per increase in sample size was attenuated at the largest sample sizes (**Supplementary Figure 7**). This suggests that once sufficiently well-powered sample sizes are reached within a given ancestry group, increasing sample sizes of other previously under-represented groups will most advance GWAS discovery.*

3. Biological interpretation. With increasing functional resolution becoming increasingly available, I was surprised not to see any downstream follow up, even if just cell type or other functional enrichment analyses for association signals.

We agree there are substantial translational opportunities with these findings, and GLGC has a long history of subsequent *in silico* and functional analyses.

[redacted]

Nevertheless, we now include an additional figure (**Figure 2**) further illustrating the improvement in fine mapping after trans-ancestry analysis in the *DMTN* region and comparison with the colocalized liver eQTL signal for *DMTN*.



4. The authors describe that polygenic scores are equally accurate in African American and European ancestry populations after a tuning strategy in African Americans that they describe as selecting SNPs in Europeans and effects in African Americans.

We apologize for the lack of clarity; this is not what we did for the main PRS. In our primary analysis, we used weights from trans-ancestry analysis after a single parameter optimization step based on multiple ancestry groups jointly analyzed in the UK Biobank.

In addition, we performed sensitivity analyses where we used the European PRS variant list and African effect sizes to generate weights, and vice-versa, to assess whether the refinement in variant list or weights was more important for PRS portability and found that both variant choice and accurate effect size estimates are important for genetic prediction in African Americans.

We have updated the text in the results section to make this more clear:

We next aimed to investigate the role of variant selection and effect size estimate variability in PRS performance between ancestry groups. We found that European ancestry GWAS-derived scores could be tuned to have improved prediction in African-American individuals by optimizing the score parameters based on testing in Africans rather than Europeans and substituting African ancestry-derived weights for

*the included variants (**Supplementary Figure 17**). However, the polygenic score derived from trans-ancestry meta-analysis was most predictive in all ancestries examined and performed similarly across most ancestry groups (adjusted $R^2 \sim 0.10-0.16$), with the trans-ancestry-derived PRS explaining an equal proportion of variance of LDL-C measurement in African American (0.16) and European-ancestry (0.16) individuals.*

This approach has been described previously and was recently shown to retain population stratification issues, for example seen previous in the GIANT study (Zaidi & Mathieson, bioRxiv following from Berg et al and Sohail et al, eLife 2019). Have the authors tested these biases, for example by looking at how effect sizes correlate with PCs in a global reference panel like 1000 Genomes and by looking across admixture fractions (if admixed samples weren't excluded per comment #1) to test trends in % variance explained with ancestry? Given the trans-ancestry nature of the study, it would be helpful to more broadly understand whether and the extent to which there is residual population stratification in GWAS summary statistics.

We thank the reviewer for this suggestion. We have now added a comparison of the trans-ancestry PRS with PCs in 1000 Genomes individuals (**Supplementary Table 16**) which showed that the significant correlations with PCs are captured by the PCs included in our prediction models. We have also added an analysis of PRS prediction by admixture fraction among African American individuals in MGI and MVP which found similar prediction of LDL irrespective of admixture proportion (**Supplementary Figure 16**). We have updated the text to include these results:

*We further assessed the prediction of polygenic scores relative to the proportion of African admixture in MGI and MVP participants. PRS prediction was similar for participants across all quartiles of African ancestry proportions (**Supplementary Figure 16**). Previous studies have suggested that population stratification may influence the predictive ability of polygenic scores across diverse populations²⁹. We tested for correlation between the trans-ancestry polygenic score and principal components (PCs) in 1000 Genomes individuals. Significant correlation was observed between the trans-ancestry polygenic score and principal components 1 and 3 only ($p\text{-value} < 0.0025$; 0.05/20 tested PCs), indicating that the minimum of 4 PCs used in the polygenic score analysis were sufficient to correct for this (**Supplementary Table 16**).*

5. Phenotype. The authors describe adjusting the phenotype by statin use (reduce LDL-C levels proportionally by a factor 0.7). Is natural variation in the presence of medications consistent with a linear reduction, particularly across different races/ethnicities? Is dose dependence expected? Are associations consistent among those on versus not on statins? Some additional citations here would be helpful. Because statins appear to not be prescribed at the same level across ethnicities, its impact on trans-ancestry analyses are important.

In the 2005 meta-analysis of 14 trials of statins (including several different dosages and types), a weighted mean reduction in LDL cholesterol at 1 year of 1.09 mmol/L was observed, which corresponds to a factor of 0.7 [range 0.6-0.9] (PMID: 16214597). We have added a citation to this meta-analysis to the manuscript. This is generally consistent with studies conducted in single ancestry groups. For example, the ARIES trial compared Rosuvastatin (10 or 20mg) vs. Atorvastatin (10 or 20mg) in African Americans and observed reductions of LDL at 6 weeks by a factor of 0.54-0.68 (PMID:16442368). The Japan Lipid Intervention Trial of simvastatin (5-10mg) observed a mean reduction in LDL at 6 months by a factor of 0.74 (PMID:12655157). This study also noted the difference in standard dosing between Japanese and western studies, but in each case a similar reduction in LDL was observed. A multi-ancestry study of LDL reduction from electronic health records in the

GERA cohort observed a median reduction in LDL by a factor of 0.66 (PMID:30354326). The absolute reduction in LDL does vary by dose and medication type (PMID:30354326). However, clinicians may prescribe medications at varying strengths in order to reach a target LDL value for an individual, rather than prescribing maximal doses which would therefore be better approximated by a single mean factor rather than medication-specific factors.

In the present analysis we chose to use the mean reduction factor rather than more complicated models involving dosages, medication types, and genetic impact on statin response, as this information is not likely to be available from all cohorts and would also depend on medication adherence. We note that inaccurate correction for statin usage would most likely reduce power, rather than result in false positive associations, and potentially reduce PRS portability, which is not observed here.

6. I have concerns about some of the underlying data. The PCs in Supp Fig 1 have an error, as PC1 separates EUR from EAS rather than AFR from all other populations, inconsistent with population genetics. Perhaps a labeling issue? Additionally, it is unclear what is shown—have most admixed populations been removed altogether? The lack of a cline on the PCs is unexpected for admixed populations described, including African American and Hispanic/Latino populations. Following from these, I have concerns about Figure 1. While tSNE plots are purely intended for visualization and have no interpretable significance, their use with genetic data in ancestrally biased cohorts are misleading. This plot is no exception as it would most intuitively suggest that European ancestry populations have more variation than all other ancestries, which is clearly not the case. PCA plots are preferable as these generate the covariates actually used in analyses and provides a more straightforward interpretation of the data, albeit with more dimensions. UMAP also performs dimensionality reduction with PCs but is less sensitive to hyperparameter tuning than tSNE, better preserving the global structure of PCs.

The PCA plots were calculated from summary statistics rather than from individual level data and so they are not expected to show the same trends as for PCA plots from individual level data. We intended to use these figures to show a summary of the ancestries and populations included in our meta-analysis and have clarified the figure legend to note that these were generated from summary level data. We have moved the tSNE plot from the main text to the supplementary information where it is presented alongside the individual PC plots. To additionally better describing contributing cohorts, we have added country of recruitment for all 280 cohorts to **Supplementary Table 1**.

Minor comments:

1. Data availability. It is critical that ancestry-specific as well as trans-ancestry GWAS summary statistics be made available. A dummy link is included but does not yet work—there should be a process to ensure these are accessible. Relatedly, it can be helpful to the field to be able to test analyses by holding out a cohort such as the widely accessed UK Biobank data for testing methods, assumptions, etc as described in the polygenic score section. Will the authors make cohort-level statistics available or possibly a set of summary statistics withholding UK Biobank upon publication?

Thank you for your interest in these results. As noted in our response to reviewer 2, we will make all single-ancestry and trans-ancestry GWAS summary statistics and PRS weights publicly available upon publication. The Global Lipids Genetics Consortium will also release

PRS weights from the GWAS excluding UKbb, for future methods development. GLGC has a long track record of summary statistic sharing, and remains committed to this.

2. It would be helpful to have a more prominent summary of number of associated loci that are trait- and ancestry-specific versus shared.

Thank you for this suggestion, we have added columns to **Supplementary Tables 3 and 6** listing the overlap between ancestries and traits.

3. Relative contributions of ancestries by sample size. Some of the explanations could be clearer. For example, in lines 779-782, the authors attribute the larger number of novel associations identified in AFR to the allele frequency differences between African and European populations. Another major contributor is the larger effective population size of the African populations (i.e. more genetic variants overall in these populations), as a population with high sample size that has been isolated could be equally differentiated but have fewer genetic variants overall.

Thank you for catching this. We have clarified this sentence to also state that this could be due to higher levels of genetic diversity among African-ancestry individuals.

*“Meta-analysis of African-ancestry individuals (N~99k, primarily African American) identified more ancestry-specific loci (15 unique to AdmAFR) than analyses of each of the other non-European ancestry groups (6 loci unique to EAS, 6 to HIS, 1 to SAS), a difference likely attributed to allele frequencies being most different between African and European ancestry populations and to African populations having a greater overall level of genetic diversity²⁷. For these variants, the frequency of the African-ancestry minor allele in European ancestry populations ranged from 0.03% to 98.9%, with a median 4.5-fold difference in allele frequencies between these groups (**Supplementary Table 4, Figure 1**).”*

4. I'm not sure how to compare the #'s of unique loci in line 779-882 with the population-specific variants that do not remain significant (lines 793-796 – shouldn't these be strictly less?).

Thank you for pointing out this section didn't read well. We have reworded this portion of the text to more accurately state that these are the variants from ancestry-specific analysis rather than ancestry-specific variants.

*“The majority (98%) of index variants from ancestry-specific analysis remained significant ($p\text{-value} < 5 \times 10^{-8}$) after meta-analysis across all ancestry groups, although 15 AdmAFR, 9 EAS, 3 HIS, and 1 SAS index variants from ancestry-specific analysis did not (trans-ancestry $p\text{-value}$ 7.7×10^{-6} to 5.9×10^{-8} , **Supplementary Figure 6**).”*

5. In lines 805-808, it could be clearer that variant effect size *estimates* can be different because of LD patterns, but that in this scenario the causal effect itself can remain the same.

Thank you! We have changed this sentence to specify the effect size estimates rather than effect size.

“Finding different variants across ancestries could be due to variation in allele frequencies -- with some variants being more common to one ancestry group -- and/or variability in effect size estimates, which could be a reflection of differing patterns of LD with the underlying causal variant or interaction with environmental risk factors.”

6. The authors describe a very large number of associated loci. What fraction of the genome do independent genome-wide significant associations cover?

Good question. We have added a sentence to the paper to describe the fraction of the genome containing significant associations of the genome.

“We find that 13% of the genome overall is associated with lipid levels, with 7%, 5%, 6%, 5%, and 6% of the genome associated with HDL-C, LDL-C, TG, nonHDL-C, and TC, respectively.”

Reviewer Reports on the First Revision:

Referee #1 (Remarks to the Author):

The authors have put in lots of effort for this revision and the additional figures and analyses add further to demonstrating the advantages to including multiple ancestries in fine-mapping and PRS.

One of my concerns remains. There is not an obvious direct link between the cited classical paper of Schwarz and the Schwarz-like approximation inferred by the authors, $BF = \exp[(X - K \cdot \ln(N))/2]$. The assumptions made by the authors to obtain their form of BF should be made clear, and also an explanation or derivation of how they reach the form $BF = \exp[(X - K \cdot \ln(N))/2]$ from Schwarz.

Referee #2 (Remarks to the Author):

In the revision the authors have addressed some of the practical concerns. The addition of African studies and demonstrations of PRS performance in these samples is a valuable addition. However, principal questions about the suitability of this manuscript for Nature remain. As summarised by reviewer 3 this is a large GWAS but has limited (none?) technological innovation or biological insights. GWAS in good field journals would usually include some follow up. The authors explain that they prefer to describe the work in three separate papers. Clearly, more than a long list of rs-ids for novel hits can be expected of a Nature-worthy publication.

In response to the question of suitability for Nature, the authors make the following statement, clarifying what they consider their key finding:

“this is the first demonstration that a multi-ancestry PRS performs equally well in continental Africans, admixed Africans *and* Europeans which is why we thought this work would be of interest to a broad audience. It is an important finding that PRS can and should be equally informative in diverse populations, and that improving the diversity of the discovery GWAS is sufficient to achieve that so as not to exacerbate racial health inequities in future approaches aimed to prevent disease using genetics to predict at-risk individuals.”

And also:

“This is the first demonstration of a PRS that works as well in African and African-American cohorts as it does in European cohorts, for any trait. We also examine differences in performance within African continental population groups. Of note, racial/ethnic differences in PRS prediction remain a key barrier to clinical use; our mitigation of this important disparity provides a roadmap for other complex diseases representing an overall conceptual advance for PRS. We provide guidance to the field on how to achieve that for other traits and diseases by leveraging a very large-sample size and a well-powered polygenic quantitative trait. We demonstrate that both the variant selection and ancestry-specific effect sizes are critical to the PRS performance, although the latter more so. We also demonstrate that studies in well-powered African-American cohorts provide strongly portable PRS scores, even to West and South African cohorts with no European admixture. This study has important implications for both cohort development and PRS methods development

across a range of traits and diseases, suggesting that existing PRS tools are suitable for prediction in other ancestries so long as sufficient sample sizes representative of diverse population groups are reached.”

The authors have done empirical PRS analyses for five closely linked lipid traits. They are suggesting that the good prediction performance of the lipid PRS in different ancestry groups is generalisable to other traits and to other groups: “existing PRS tools are suitable for prediction in other ancestries so long as sufficient sample sizes representative of diverse population groups are reached”. And therefore this represents an “overall conceptual advance for PRS”. It is entirely inappropriate to make such a strong claim. The authors have not done simulations or considered any other traits. Complex traits differ in their genetic architecture. Others have found differences between traits (<https://www.nature.com/articles/s41588-019-0379-x>) and trans-ancestry genetic correlations for different traits range from low to very high (<https://doi.org/10.1002/gepi.22173>).

Referee #4 (Remarks to the Author):

The authors detail an incredible amount of work to assemble the largest lipids genetics consortium and detail their work in discovery, fine-mapping, and PRS development. Congratulations on this monumental effort! I have provided some major and minor points below. In general, it is unclear what the focus and narrative of the paper is supposed to be, whether it is focused on the large collaborative effort (for which more comparisons to prior publications should be made) or focused on the value of multiethnic cohorts (for which more apples-to-apples comparisons should be made where ancestry is the key factors). I think both could be great and meaningful.

MAJOR POINTS

- There is a lot jammed into a relatively short article and I am sympathetic to the pressure to keep things short. You all did a ton of work and all of it is interesting! However, as a fresh set of eyes it was really difficult to keep things straight. I would suggest going through and provide a few more statements setting up each of the experiments/analyses, as well as giving a brief statement for what they would show. This is already done somewhat, but the flow of the article would benefit from editing with this in mind.
- I would be interested to know if there are differences in trends between the traits and between populations besides AdmAFR and EUR. Currently many of the analyses pool all traits together when reporting trends, but given the trans-ethnic nature of the paper, it would be nice to see (even with just a sentence or two) if these trends are trait-specific within lipids.
- For the LDL-C PRS, the authors used two different methods (PRS-CS, P+T) with several thresholds for P+T. The authors state (page 27, lines 960-961) that they “optimized the scores”, but it is unclear how the scores were optimized, or what the optimal scores in fact were. There is no mention of whether PRS-CS or P+T are better (or the parameters). This information is only found by going through the supplement. In general, I found this section a bit confusing as the authors compare scores that differ in (a) ancestry and (b) sample size, as well as (c) methods and tuning parameters (which depend on a and b).
- The authors state that the PRS derived from ALL have better performance than those from stratified GWAS. However, this is dependent on the PRS development method per Supp Figure 15. It would be helpful to add some text in saying for a specific method (PRS-CS or P+T with specific thresholds), then the trends hold. It isn’t necessarily a useful comparison to say that the average score over a range of different methods and thresholds, this PRS performs best. In reality, you would choose an optimal score and move forward with that one score, not use an entire suite of PRS and take the mean to give back an individual’s risk.
- To investigate the role of sample size, the authors look at the performance of their LDL-C PRS using smaller subsets of EUR with various combinations of AdmAFR or non-EUR samples. I appreciate the point about how it doesn’t matter how large the EUR sample sizes get, they can’t make up ground in non-EUR populations (specifically AFR). However, when comparing the performance of the PRS from ALL vs. stratified analyses, the elephant in the room are the vastly

different sample sizes between populations. This is shown in Supp Figure 15. To address these differences, a comparison of sample sizes would compare stratified analyses of the same size (downsampled in this case) compared to a transethnic analysis with total sample size the same as any stratified analysis. That way the only thing you are varying is the ancestry of participants, not the sample size.

- Lines 838-840 describe the proportion of variance explained by trans-ancestry index variants by trait. Do these differ by ancestry?
- The authors use Popcorn to estimate genetic correlations. However, Popcorn is not well-suited for admixed populations due to long-range LD. How did the authors address this limitation and possible consequences in their data?
- Many of the trends described in the first few sections collapse all traits. Do the trends in finemapping and discovery between ancestries differ by trait? Are some traits more ancestry-specific?

MINOR POINTS

- The authors say in the first line of the PRS section that "We next evaluated the potential of polygenic scores to predict individuals at risk of CAD through elevated LDL-C in blood." However, I saw no extension of the PRS for LDL-C in the prediction of CAD. Did they mean to say that they chose LDL-C to look at the PRS because it would have clinical relevance, as it is a risk factor for CAD?
- I can't read the numbers in the blue boxes in Supplementary Figure 10. Either make them in white or choose a different color/transparency.
- Supplemental Figure 7B needs an x-axis labels.
- In addition to the total sample sizes in Table 1, it would be helpful to also have the number of studies contributing that ancestry and the mean (and range) of their sample sizes. As this is the largest meta-analysis to date, it would also be helpful to know what the largest published GWAS for these traits in these ancestries was (this would show how impressive this effort is compared to prior studies).
- Figure 1 compares allele frequencies across groups. However, it would be helpful to have a plot that looks at the difference in allele frequencies between populations in a pairwise fashion. For example, are these variants that are at 2% in one population and 80% in another? Or are the differences on average smaller?

Here are how I think R3's comments were addressed:

1. R3's first comment asked for comparisons of the PRS performance using this latest round of transethnic meta-analysis results compared to previous publications. Specifically, they wanted to see if there was better performance using the summary statistics from this publication or from previous publications for the Admixed African group. The authors did not do this analysis. They instead looked at the different sample sizes of Europeans and the trends in performance in AdmAFR, which is interesting, but not the point.
2. R3 wanted to see discovery by different down sampled EUR subsets to truly get at the trend in sample size vs variant discovery. The authors have added these analyses to both the supplement and main text.
3. R3 was disappointed to not see any functional follow-up, which would be needed to assess how fine-mapping helps understand biology. The authors state that this is for other manuscripts, which is disappointing considering this is in Nature and usually the scope of results is not as narrow, but understandable given the large consortium and competing interests.
4. R3 had comments about the tuning procedure, which was clarified somewhat by the authors for the PRS section. Honestly I had a hard time reading this section. There are a lot of things going on and it is unclear what the final message is. R3 also wanted to know the relationship between admixture proportions and risk score. The authors stratified by quintiles of ancestry and included in the analysis. It would have been nice to see a bit more about this, but I understand that real estate is limited.

5. R3 had concerns about the phenotypes, which were explained further with references.
6. Confusion around a figure was addressed and the concerning figure was moved to the supplement. I agree with both the original concern and the response from the authors.

All minor points seem to be addressed sufficiently.

Author Rebuttals to First Revision:

Referee #1 (Remarks to the Author):

The authors have put in lots of effort for this revision and the additional figures and analyses add further to demonstrating the advantages to including multiple ancestries in fine-mapping and PRS.

Thank you for your review.

One of my concerns remains. There is not an obvious direct link between the cited classical paper of Schwarz and the Schwarz-like approximation inferred by the authors, $BF = \exp[(X - K \ln(N))/2]$.

The assumptions made by the authors to obtain their form of BF should be made clear, and also an explanation or derivation of how they reach the form $BF = \exp[(X - K \ln(N))/2]$ from Schwarz.

We apologize for the lack of clarity related to the derivation of the approximate Bayes' factor, which is based on the Schwarz Criterion (Schwarz, 1978), and is reviewed by Kass & Raftery (1995). We have included the following derivation in the Supplementary Information. This approach has the advantage that it avoids the introduction of prior densities of the model parameters.

Consider two models, M_0 and M_1 . Let $\hat{\theta}_k$ denote the maximum likelihood estimates of model parameters under model M_k and let d_k denote the dimension of model M_k . The Schwarz Criterion is given by

$$S = \log f(y|\hat{\theta}_1, M_1) - \log f(y|\hat{\theta}_0, M_0) - \frac{(d_1 - d_0)}{2} \log(n)$$

where y are observed data and n is the sample size. The Bayes' factor in favour of model M_1 over M_0 is then approximated by $\exp(S)$.

In the context of our study, the null model M_0 corresponds to allelic effect sizes fixed at 0, whilst under the alternative model M_1 allelic effect sizes are unconstrained. The difference in log-likelihoods between the two models is given by $X^2/2$, where X^2 is the deviance between the two models, which is approximated here by the observed chi-square statistic,

$$X^2 = \frac{\beta^2}{SE^2}$$

It then follows that

$$BF \approx \exp\left[\frac{X^2 - \log(n)}{2}\right]$$

References:

Kass RE, Raftery AE (1995). Bayes factors. *Journal of the American Statistical Association*, 90: 773-795.
Schwarz G (1978). Estimating the dimension of a model. *Annals of Statistics*, 6: 461-464.

Referee #2 (Remarks to the Author):

In the revision the authors have addressed some of the practical concerns. The addition of African studies and demonstrations of PRS performance in these samples is a valuable addition. However, principal questions about the suitability of this manuscript for Nature remain. As summarised by reviewer 3 this is a large GWAS but has limited (none?) technological innovation or biological insights.

We have examined the role of ancestry in genetic discovery in an unprecedented sample size using state-of-the-art genetic discovery approaches (many of which we developed). Through a comprehensive examination of the portability of polygenic scores in hundreds of thousands of individuals from diverse continental groups including people from Africa, East Asia and African Americans, we have documented the value of a diverse genetic ancestry in the development of prediction tools (polygenic scores) which holds promise to prevent atherosclerotic disease across population groups without exacerbating existing health inequalities. We feel reasonably confident that our observations for PRS (that trans-ancestry PRS will perform the best in diverse sets of cohorts) will also apply to many other complex traits even if it doesn't apply to all other traits. We also provide examples of coding variants associated with a variety of important diseases as novel drug targets, loci that have been fine-mapped to a single functional variant through access to a large diverse sample set. A deep examination of the biological implications of 2,727 variants including biological pathways, candidate loci, eQTLs, PheWAS, sex-specific loci and X chromosome associations unfortunately does not fit into a Nature-length article in addition to the multi-ancestry topics we have already included.

GWAS in good field journals would usually include some follow up. The authors explain that they prefer to describe the work in three separate papers. Clearly, more than a long list of rs-ids for novel hits can be expected of a Nature-worthy publication.

We agree that GWAS in good field journals often includes follow-up (see our previous GWAS studies of atrial fibrillation in Nielsen et al *Nat Genet* 2018, TSH in Zhou et al *Nat Comm* 2019, TM6SF2 in Holmen et al *Nat Genet* 2014, blood lipids in Klarin et al *Nat Genet* 2018 as evidence of this well-used approach). However, this manuscript extends beyond a typical GWAS that would drill down into new associated variants, biological implications at specific loci and single locus findings. Instead, we focus and address the bigger scientific question of whether studying non-European ancestries is worth the additional effort to expand recruitment into new sites and new demographics. We evaluated this in a scientific (and quantitative) way to demonstrate that by increasing diversity, improvements are likely to be made in fine-mapping and portability of polygenic prediction, as well as modest improvements in the number of discovered loci or associated variants. We added additional language to clarify these conclusions.

In response to the question of suitability for Nature, the authors make the following

statement, clarifying what they consider their key finding:

“this is the first demonstration that a multi-ancestry PRS performs equally well in continental Africans, admixed Africans *and* Europeans which is why we thought this work would be of interest to a broad audience. It is an important finding that PRS can and should be equally informative in diverse populations, and that improving the diversity of the discovery GWAS is sufficient to achieve that so as not to exacerbate racial health inequities in future approaches aimed to prevent disease using genetics to predict at-risk individuals.”

And also:

“This is the first demonstration of a PRS that works as well in African and African-American cohorts as it does in European cohorts, for any trait. We also examine differences in performance within African continental population groups. Of note, racial/ethnic differences in PRS prediction remain a key barrier to clinical use; our mitigation of this important disparity provides a roadmap for other complex diseases representing an overall conceptual advance for PRS. We provide guidance to the field on how to achieve that for other traits and diseases by leveraging a very large-sample size and a well-powered polygenic quantitative trait. We demonstrate that both the variant selection and ancestry-specific effect sizes are critical to the PRS performance, although the latter more so. We also demonstrate that studies in well-powered African-American cohorts provide strongly portable PRS scores, even to West and South African cohorts with no European admixture.

This study has important implications for both cohort development and PRS methods development across a range of traits and diseases, suggesting that existing PRS tools are suitable for prediction in other ancestries so long as sufficient sample sizes representative of diverse population groups are reached.”

The authors have done empirical PRS analyses for five closely linked lipid traits. They are suggesting that the good prediction performance of the lipid PRS in different ancestry groups is generalisable to other traits and to other groups: “existing PRS tools are suitable for prediction in other ancestries so long as sufficient sample sizes representative of diverse population groups are reached”. And therefore this represents an “overall conceptual advance for PRS”. It is entirely inappropriate to make such a strong claim. The authors have not done simulations or considered any other traits. Complex traits differ in their genetic architecture. Others have found differences between traits (<https://www.nature.com/articles/s41588-019-0379-x>) and trans-ancestry genetic correlations for different traits range for low to very high (<https://doi.org/10.1002/gepi.22173>).

We thank the reviewer for this insightful comment, and the important feedback that the claim as currently written may be too far-reaching. We agree and have minimized the claim of generalizability to other diseases and traits. We acknowledge that further study is warranted to validate the portability of trans-ancestry risk scores across diverse traits with varying genetic architectures. We have revised this portion of the manuscript to explicitly state that additional studies are needed to confirm these findings in other traits.

We emphasize the need to study other diseases and traits to assess the generalizability of the finding that a trans-ancestry polygenic score with ancestry-specific normalization is most informative. While many traits show a high degree of shared genetic correlation across ancestries^{36,42,43} others have distinct genetic variants with large effects that are more common in specific ancestry groups⁴⁰ which may limit the transferability of trans-ancestry scores for other phenotypes.

However, we can't expect to understand the portability of all genetic risk scores for all diseases and traits in exceptionally well-powered sample sizes of this magnitude right now. The answer to this question would take the combined power of all cohorts with genetic data and disease/trait consortia working together for years. In our opinion, a demonstration of multi-ancestry performance of polygenic scores and GWAS discovery for lipid traits is an excellent starting place for the field to jump off from and to encourage more diverse studies in other diseases and traits. We would like to apprise the field, based on findings from this large multi-ancestry study of lipids, appropriate expectations of where gains are likely to be seen (fine-mapping, PRS portability and some limited ancestry-specific variants) and where the gains from non-European ancestry GWAS may be more limited (novel locus discovery).

We believe we present the results in an unbiased and data-driven manner.

Referee #4 (Remarks to the Author):

The authors detail an incredible amount of work to assemble the largest lipids genetics consortium and detail their work in discovery, fine-mapping, and PRS development. Congratulations on this monumental effort! I have provided some major and minor points below. In general, it is unclear what the focus and narrative of the paper is supposed to be, whether it is focused on the large collaborative effort (for which more comparisons to prior publications should be made) or

focused on the value of multiethnic cohorts (for which more apples-to-apples comparisons should be made where ancestry is the key factors). I think both could be great and meaningful.

Thank you for your review and suggestions to improve our manuscript. We have revised the manuscript to focus on the value of diverse ancestry groups (the latter part of your description). We tried to improve the apples-to-apples comparisons, as you suggest, by adding new analyses focused on portability of the PRS derived from different ancestry proportions, but keeping constant the sample size (N~100k) and PRS methodology (comparison within PRS-CS and a separate comparison using optimized pruning & thresholding approaches).

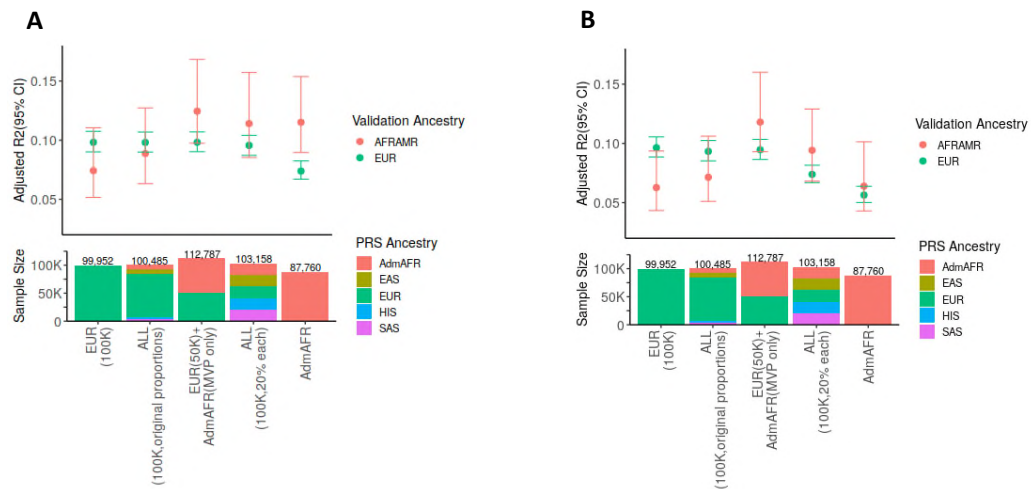
We have added the following new figure to visualize these results:

Supplementary Figure 17: Comparison of PRS source ancestry and sample size with prediction in European and African-American individuals.

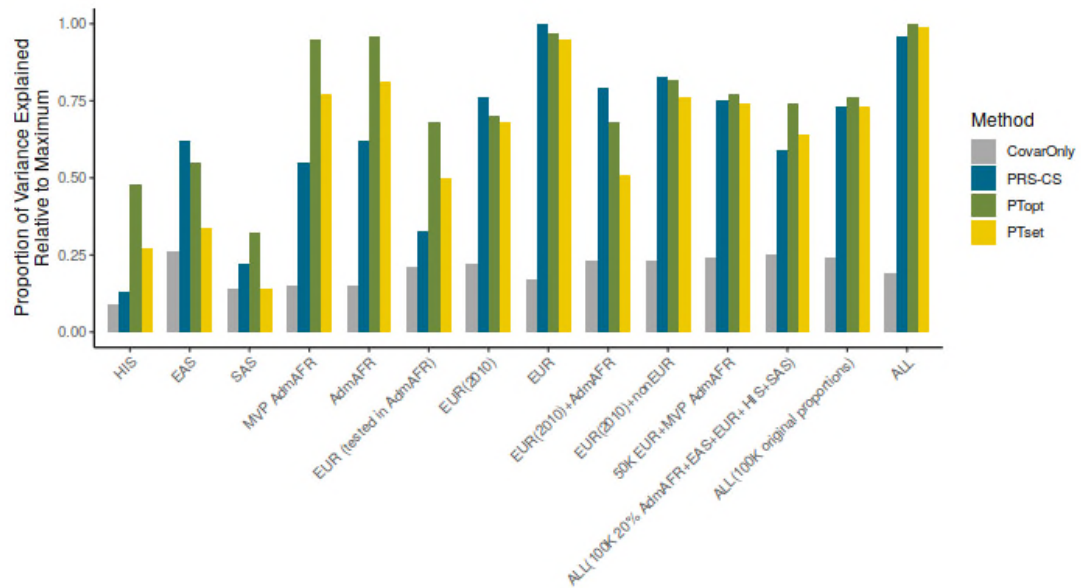
Error bars represent 95% confidence intervals.

A) At constant sample size (~100,000) and using only pruning and thresholding to create risk scores used for prediction in the Michigan Genomics Initiative (MGI)

B) At constant sample size (~100,000) and using only PRS-CS to create risk scores used for prediction in the Michigan Genomics Initiative (MGI)



We also now add an explicit comparison of three PRS-generating methods in the optimizing dataset for each ancestry: PRS-CS, pruning & thresholding optimized, and pre-set generic pruning & thresholding parameters (P of 5×10^{-5} , r^2 0.1 and distance 500kb) – new Supplementary Table 15 and Supplementary Figure 15.



MAJOR POINTS

- There is a lot jammed into a relatively short article and I am sympathetic to the pressure to

keep things short. You all did a ton of work and all of it is interesting! However, as a fresh set of eyes it was really difficult to keep things straight. I would suggest going through and provide a few more statements setting up each of the experiments/analyses, as well as giving a brief statement for what they would show. This is already done somewhat, but the flow of the article would benefit from editing with this in mind.

Thank you for this helpful comment. Several senior members of the writing team have now gone through the manuscript to further improve readability and flow. We have extensively revised the entire manuscript. We hope it is more easily understood to a broad audience as well as field-specific experts now.

- I would be interested to know if there are differences in trends between the traits and between populations besides AdmAFR and EUR. Currently many of the analyses pool all traits together when reporting trends, but given the trans-ethnic nature of the paper, it would be nice to see (even with just a sentence or two) if these trends are trait-specific within lipids.

We thank the reviewer for this suggestion and have added a summary of the trait differences for ancestry-specific locus discovery. We have examined these analyses separately by trait for each non-European ancestry-specific locus. The six Hispanic-specific index variants were on chromosome 11 and are now described in the main text. The one South Asian-specific locus was associated with TG, while the East Asian-specific loci included four HDL-C loci and two LDL-C/nonHDL-C/TC loci. The Admixed African-specific loci included two loci associated with HDL-C/LDL-C/TC, three loci specific to HDL, and ten loci associated with either LDL-C, nonHDL-C, and/or TC. The effect sizes of variants identified from each ancestry-specific analysis included both common variants of small effect and rare variants with moderate to large effect (Supplementary Table 3 and Supplementary Figure 3).

We have now added the overlap between traits and ancestries for all index variants in **Supplementary Table 3**.

The following text has been added to the manuscript:

*We next assessed whether the ancestry-specific variants were enriched by trait or ancestry. The six Hispanic-specific index variants were all associated with triglycerides and concentrated within a region in length of several megabases on chromosome 11 that has been previously identified to be associated with high TG levels among individuals with Indigenous American ancestry^{31,32}. Other ancestry-specific associated variants were relatively evenly distributed among the different lipid traits (**Supplementary Table 3**).*

- For the LDL-C PRS, the authors used two different methods (PRS-CS, P+T) with several thresholds for P+T. The authors state (page 27, lines 960-961) that they “optimized the scores”, but it is unclear how the scores were optimized, or what the optimal scores in fact were. There is no mention of whether PRS-CS or P+T are better (or the parameters). This information is only found by going through the supplement. In general, I found this section a bit confusing as the authors compare scores that differ in (a) ancestry and (b) sample size, as well as (c) methods and tuning parameters (which depend on a and b).

We apologize that this was confusing as previously written. We have revised the manuscript text (see below) to directly describe the PRS optimization approach, as we agree this is likely to be of interest to readers. In addition, we have added a column to **Supplementary Table 14** that lists the single best risk score method and parameters for each ancestry that performed best in the testing dataset and was taken forward for use in the validation cohorts.

New text added:

*We evaluated the potential of polygenic scores to predict elevated LDL-C, a major causal risk factor of CAD, in diverse ancestry groups. To optimize and test polygenic scores, we created three datasets with no overlapping individuals. We used these datasets to: i) perform either ancestry-specific or trans-ancestry GWAS to estimate variant effect sizes, ii) optimize risk score parameters, and iii) evaluate the utility of the score. For each GWAS (either trans-ancestry or ancestry-specific), we created multiple polygenic scores using: 1) a subset of independently associated variants (pruning and thresholding) or 2) a genome-wide approach (PRS-CS³⁹). We next tested all available scores in the optimizing dataset, which was matched for ancestry to the GWAS (AdmAFR, EAS, EUR, SAS, ALL from UK biobank or HIS from Michigan Genomics Initiative, **Supplementary Figures 14-16, Supplementary Tables 14-16**). We took the top-performing score from each GWAS dataset forward for independent evaluation in separate datasets: PRS-CS for East Asian, European, and European 2010⁵ scores, and an optimized pruning and threshold-based score for all other GWAS datasets. We evaluated the PRS in 8 cohorts of individuals (N=295,577, **Supplementary Table 17**) not included in the discovery GWAS, from 6 ancestral groups: East Asian (146,477), European American (85,571), African American (21,730), African (2,452 East Africa, 4,972 South Africa, 7,309 West Africa), South Asian (15,242), Hispanic American (7,669), and Asian American (4,155).*

- The authors state that the PRS derived from ALL have better performance than those from stratified GWAS. However, this is dependent on the PRS development method per Supp Figure 15. It would be helpful to add some text in saying for a specific method (PRS-CS or P+T with specific thresholds), then the trends hold. It isn't necessarily a useful comparison to say that the average score over a range of different methods and thresholds, this PRS performs best. In reality, you would choose an optimal score and move forward with that one score, not use an entire suite of PRS and take the mean to give back an individual's risk.

We apologize for the confusion and provide clarification. We selected a single optimal score for the trans-ancestry analysis and a single score for each of the ancestry-specific analyses, but did not average the results across multiple scores. We instead compare each type of score individually. For example, if we were trying to predict lipids in African-American ancestry individuals, how would a score derived from a European-only GWAS compare to a score from a multi-ancestry GWAS. We did not average any scores in a single individual.

We have now repeated the comparisons while keeping the PRS optimization approach constant. We added **Supplementary Figure 17a** which shows using a single method (pruning and thresholding only) at a constant sample size of ~100,000 individuals the improved transferability of the trans-ancestry derived score relative to either the Admixed

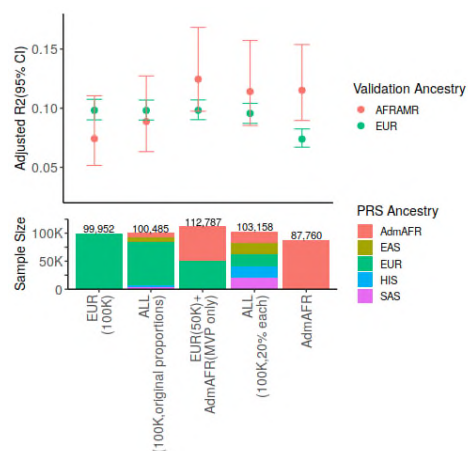
African or European-specific scores. Similarly, in Supplementary Figure 17b, we compare the transferability of a score generated using PRS-CS from different GWAS.

- To investigate the role of sample size, the authors look at the performance of their LDL-C PRS using smaller subsets of EUR with various combinations of AdmAFR or non-EUR samples. I appreciate the point about how it doesn't matter how large the EUR sample sizes get, they can't make up ground in non-EUR populations (specifically AFR). However, when comparing the performance of the PRS from ALL vs. stratified analyses, the elephant in the room are the vastly different sample sizes between populations. This is shown in Supp Figure 15. To address these differences, a comparison of sample sizes would compare stratified analyses of the same size (downsampled in this case) compared to a transethnic analysis with total sample size the same as any stratified analysis. That way the only thing you are varying is the ancestry of participants, not the sample size.

We thank the reviewer for this suggestion. We have now added an additional set of analyses testing ~100,000 individual subsets of Admixed African, European, Admixed African + European, or ALL ancestries in either equal proportions (20% each) or matched to the proportions of the full 1.65 million individual GWAS (predominantly European, with smaller proportions of non-European ancestries). These analyses were performed using only a pruning and thresholding approach rather than selecting between PRS-CS and pruning and thresholding. As shown in **Supplementary Figure 17a**, we find that the trends observed in the full dataset remain. Ancestry-mismatched polygenic scores are worse predictors of LDL cholesterol than ancestry-matched scores, and trans-ancestry derived polygenic scores show better transferability to other ancestries.

Please see the newly added text to the manuscript describing these results and the relevant figure (**Supplementary Figure 17a**):

*We first evaluated the role of GWAS diversity on polygenic score accuracy. We fixed the sample sizes at ~100k and created five different GWAS: EUR and AdmAFR single-ancestry GWAS, a half EUR, half AdmAFR bi-ancestry GWAS, and two trans-ancestry GWAS, one with equal numbers of the five ancestries and one where each ancestry matched the proportion in the full 1.65m meta-analysis. Using a pruning and thresholding approach, we created optimized PRS weights for each of these five different GWAS meta-analyses. The ancestry matched single-ancestry scores were similar or slightly worse predictors of LDL-C compared to the trans-ancestry scores (**Supplementary Figure 17A**). The trans-ancestry score with equal proportions of each ancestry group predicted LDL-C better among African Americans, and both trans-ancestry scores predicted LDL-C similarly well among Europeans. Lastly, the ancestry-mismatched scores predicted LDL-C less well in African Americans (65% of trans-ancestry PRS) than the ancestry-mismatched score predicted LDL-C in European Americans (77% of trans-ancestry PRS).*



- Lines 838-840 describe the proportion of variance explained by trans-ancestry index variants by trait. Do these differ by ancestry?

We thank the reviewer for this suggestion. We find that the trans-ancestry index variants explain the largest proportion of variance in Europeans, assuming a shared effect size across ancestries. The variance explained in other ancestries is lower, likely due to the large European sample size in the present meta-analysis. This further emphasizes the need to focus on recruiting individuals from other ancestry groups for future studies. We have added the following manuscript text to describe this:

*Overall, the novel trans-ancestry index variants reaching genome wide significance explained ~0.8% of the variance in each trait, with all variants explaining 12%, 13%, 9%, 13%, and 12% of the variance across all ancestries, for HDL-C, LDL-C, TG, nonHDL-C, and TC, respectively. Assuming a shared effect size across ancestries with varying allele frequencies, we find that the proportion of variance explained by the trans-ancestry index variants within each ancestry is 14%, 32%, 3%, 16%, and 20% lower on average in Admixed Africans, East Asians, Europeans, Hispanics, and South Asians, respectively (**Supplementary Table 7**).*

- The authors use Popcorn to estimate genetic correlations. However, Popcorn is not well-suited for admixed populations due to long-range LD. How did the authors address this limitation and possible consequences in their data?

Thank you for the suggestion. We have now added estimates of the genetic correlation calculated within the UK Biobank and the Million Veteran Program using individual level genetic data analyzed with the bivariate GREML approach implemented in GCTA. This method has been previously applied to calculate genetic correlations between European and African individuals in the UK Biobank for BMI and height only [PMID: 33664403]. We find that the genetic correlations calculated with either approach are similar in magnitude. The additional results are included Supplementary Table 8 and in the revised manuscript.

We tested for genome-level differences in both the genetic effect correlation and the genetic impact correlation, i.e. the genetic effect after accounting for allele frequency differences between populations, using Popcorn³⁴. Only African-ancestry effect sizes

*showed a significantly decreased correlation with other ancestry groups (p -value < 0.05 for correlation not equal to 1), with genetic impact correlations ranging from 0.49-0.80 for TC, 0.52-0.77 for LDL-C, 0.68-0.90 for TG, and 0.54-1 for nonHDL-C (**Supplementary Figures 9 and 10**). HDL-C effect sizes were highly correlated and not significantly different from 1. We additionally validated the observed differences in genetic correlation between Admixed Africans and Europeans in the UK Biobank and Million Veteran Program using bivariate GREML^{35,36} as the Popcorn method does not account for long-range LD in admixed populations. Genetic correlations between AdmAFR and EUR for HDL-C were not significantly different from 1 in the UK Biobank (possibly due to relatively small numbers of African ancestry individuals), while correlations for the other traits ranged from 0.52-0.6 in UK Biobank and 0.47-0.69 in MVP (**Supplementary Table 9**).*

- Many of the trends described in the first few sections collapse all traits. Do the trends in finemapping and discovery between ancestries differ by trait? Are some traits more ancestry-specific?

We focused on the minimal credible set for each shared association signal from any trait as many of the associations are shared across traits and because traits differ in available sample size. We do not expect any differences by trait in reduction of credible set size. However, we have added a further discussion of the trends in discovery between traits to the revised manuscript (see below).

*We next assessed whether the ancestry-specific variants were enriched by trait or ancestry. The six Hispanic-specific index variants were all associated with triglycerides and concentrated within a region in length of several megabases on chromosome 11 that has been previously identified to be associated with high TG levels among individuals with Indigenous American ancestry^{31,32}. Other ancestry-specific associated variants were relatively evenly distributed among the different lipid traits (**Supplementary Table 3**).*

MINOR POINTS

- The authors say in the first line of the PRS section that “We next evaluated the potential of polygenic scores to predict individuals at risk of CAD through elevated LDL-C in blood.” However, I saw no extension of the PRS for LDL-C in the prediction of CAD. Did they mean to say that they chose LDL-C to look at the PRS because it would have clinical relevance, as it is a risk factor for CAD?

We thank the reviewer for noticing this inadvertent error and we have revised the sentence in the manuscript text (see below)

We next evaluated the potential of polygenic scores to predict elevated LDL-C, a major risk factor of CAD, in diverse ancestry groups.

- I can't read the numbers in the blue boxes in Supplementary Figure 10. Either make them in white or choose a different color/transparency.

Thank you for reviewing the Supplementary Figures. We are grateful for your time and comments. We have updated **Supplementary Figure 10** (and also **Supplementary Figure 9**) to a lighter color scheme.

- Supplemental Figure 7B needs an x-axis labels.

We have added additional labels to **Supplementary Figure 7B**

- In addition to the total sample sizes in Table 1, it would be helpful to also have the number of studies contributing that ancestry and the mean (and range) of their sample sizes. As this is the largest meta-analysis to date, it would also be helpful to know what the largest published GWAS for these traits in these ancestries was (this would show how impressive this effort is compared to prior studies).

We thank the reviewer for this suggestion and have edited **Table 1** to include contributing cohorts and their sample size.

Table 1: Meta-analysis sample size by ancestry group

Ancestry Group	Sample Size	Number of Cohorts	Mean Sample Size per Cohort (range)	Number of Variants
European	1,320,016	123	10,928 (173-389,344)	75 M
East Asian	146,492	30	7,448 (150-131,050)	27 M
Admixed African/ African	99,432	19	5,330 (473-62,022)	48 M
Hispanic	48,057	8	6,032 (1,496-22,302)	37 M
South Asian	40,963	6	6,413 (1,796-16,110)	37 M
Total	1,654,960	164		91 M

The present meta-analysis represents a 6-fold overall increase in sample size relative to the most recent 2018 Million Veteran Program blood lipid meta-analysis¹⁴, with a 2-fold increase in sample size of Admixed African and Hispanic individuals.

*Note these studies (HRC and 1000G imputation) are independent from GLGC 2017 (exome array) and 2013 (metabohip) meta-analyses

The most recent GWAS publication of lipids (Klarin et al Nat Genet 2018) had the following sample sizes and **did not evaluate polygenic risk in non-European individuals at all**.

Ancestry Group (Klarin et al 2018)	Sample Size	Number of Cohorts
European	~216,000	1
East Asian	0	0
Admixed African/African	~57,000	1
Hispanic	~24,000	1
South Asian	0	0
Total Discovery	~297,000	1
Replication GLGC exome array Lu et al NG 2017	~319,000	73
Replication GLGC metabochip Willer et al NG 2013	~189,000	60

We have consulted with the lead junior and senior authors of the Klarin paper (Klarin and Assimes) who are also authors on this manuscript, and confirm that one principal reason they did not evaluate polygenic risk across ancestries is because large African American and Hispanic samples sizes completely independent of all discovery cohorts listed above were not available. This issue was resolved with the second large tranche of genetic data in the Million Veteran Program made available in 2018 and incorporated into this manuscript.

- Figure 1 compares allele frequencies across groups. However, it would be helpful to have a plot that looks at the difference in allele frequencies between populations in a pairwise fashion. For example, are these variants that are at 2% in one population and 80% in another? Or are the differences on average smaller?

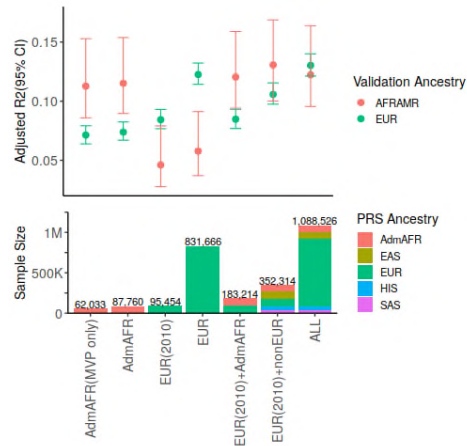
We have added additional scatter plots to **Supplementary Figure 3** comparing the allele frequencies for genome-wide significant associated variants from ancestry-specific meta-analysis vs all other ancestries in a pairwise format. Additionally, all allele frequencies across ancestry groups for each index variant are given in **Supplementary Table 4**. The vast majority of variants are common in both ancestries, though with variable frequency (eg. 40% frequency in Europeans and 70% frequency in Admixed Africans), however some of the ancestry specific variants are much more frequent in a single ancestry group (as expected). For example, the *CREB3L2* association with total cholesterol was detected in Admixed Africans due the much higher minor allele frequency of 10% in AdmAFR relative to the 0.2% frequency in Europeans despite the much larger European sample size and similar effect sizes in both ancestry groups. All data is available in Supplementary Table 4 for readers to review.

Here are how I think R3's comments were addressed:

1. R3's first comment asked for comparisons of the PRS performance using this latest round of transethnic meta-analysis results compared to previous publications. Specifically, they wanted to see if there was better performance using the summary statistics from this publication or from previous publications for the Admixed African group. The authors did not

do this analysis. They instead looked at the different sample sizes of Europeans and the trends in performance in AdmAFR, which is interesting, but not the point.

We have now added a comparison with the polygenic score developed from the full Admixed African GWAS to one developed from the MVP African American GWAS alone. We find that the increase in Admixed African sample size leads to a small increase in the prediction ability of model, as measured by adjusted R^2 and shown in **Supplementary Figure 17** and **Supplementary Table 17**. However, please note polygenic scores derived from the meta-analysis of MVP in Klarin et al 2018 were not generated or evaluated as part of that manuscript, neither from a multi-ancestry or single ancestry cohort (as noted in more detail



above).

2. R3 wanted to see discovery by different down sampled EUR subsets to truly get at the trend in sample size vs variant discovery. The authors have added these analyses to both the supplement and main text.

Thank you for acknowledging this suggested revision has already been addressed.

3. R3 was disappointed to not see any functional follow-up, which would be needed to assess how fine-mapping helps understand biology. The authors state that this is for other manuscripts, which is disappointing considering this is in Nature and usually the scope of results is not as narrow, but understandable given the large consortium and competing interests.

We concur with R4 and the editor that functional follow-up would make this manuscript unfocused and detract from the trans-ancestry focus.

4. R3 had comments about the tuning procedure, which was clarified somewhat by the authors for the PRS section. Honestly I had a hard time reading this section. There are a lot of things going on and it is unclear what the final message is. R3 also wanted to know the relationship between admixture proportions and risk score. The authors stratified by quintiles

of ancestry and included in the analysis. It would have been nice to see a bit more about this, but I understand that real estate is limited.

We have rewritten the text in this section and hope it is now more clear. On the previous revision, we had also added evaluation of the PRS by proportion of European admixture in African Americans.

5. R3 had concerns about the phenotypes, which were explained further with references.

Thank you for acknowledging this suggested revision has already been addressed.

6. Confusion around a figure was addressed and the concerning figure was moved to the supplement. I agree with both the original concern and the response from the authors.

Thank you for acknowledging this suggested revision has already been addressed.

All minor points seem to be addressed sufficiently.

We thank all four previous reviewers for undoubtedly making this a stronger manuscript by their volunteer service of reviewing of this work. We are very grateful.

Reviewer Reports on the Second Revision:

Referee #1 (Remarks to the Author):

I have no further comments.

The form of the BF now includes $\log(n)$ and a sufficient explanation is given.

Referee #4 (Remarks to the Author):

The authors have made substantial revisions to address the reviewers comments. They have focused on trans-ethnic analyses of lipids with respect to GWAS discovery, fine-mapping, and polygenic risk score development. Overall, the manuscript, specifically the results section, is very long and contains information that is not mentioned in the discussion or provided with an interpretation. Some of this can be cut out of the main text and left in the supplement. The manuscript could be tightened up some and distilled to its core points, but the necessity for this depends on the word count and formatting. I have listed specific concerns below.

1. The authors conduct an analysis looking at the number of significant loci as a function of sample size, showing a linear trend. However, statistical power is not linear with respect to sample size, but exponential. Expectation should therefore fit an exponential, not linear, trend.
2. Figure 1F shows the number of ancestries with power >80%, keeping sample size constant. However, it seems that the index variants were selected on the stratified analyses, which have much different sample sizes. The authors have eliminated one aspect of difference (sample size for power calculations), but not the vast differences in discovery sample sizes that would confound
3. On pg 22 the authors assume a shared effect size across ancestries to estimate the proportion of variance explained. It is unclear why this assumption is made, considering that on the next page they show decreased correlation between ancestries with 7% reaching a P-value threshold for heterogeneity in effects. Why weren't the variants identified from the trans-ethnic analysis, but the population-specific weights used for this calculation?
4. I found the last paragraph of the fine-mapping section confusing and unsupported with evidence

(pg 26). Given that there is no functional follow-up to identify causal alleles, what evidence do the authors have that the indel within ANGPTL3 is the more likely causal variant compared to DOCK7? Additionally, do the variants identified as having a high posterior probability in African populations have higher allele frequencies than in other populations? Or is the LD block broken up more? I would also be hesitant to say that these are ancestry-specific to Africa, given the authors only compare Africa to Europe. Therefore it may be a non-European-specific trend.

5. The authors examine the relationship between their trans-ancestry score and the principal components in 1000Genomes. It is unclear why they did not do this in their data that includes outcome data to investigate if this relationship is relevant to prediction or a possible source of bias.

6. I found the second to last paragraph of the PRS section confusing. The authors offer solutions of how to improve PRS performance when starting with EUR-derived scores. However, their suggestions seem to be to either re-select variants based on parameters in your target population and/or use weights from your target population. This seems like they are saying to basically create a new score in this new population, but I could be misunderstanding this.

7. While the authors have added supplemental materials that address some of this concern, the message is still a bit muddled when it comes to PRS performance, given the different methods (P+T, PRS-CS) and different parameters (r^2 , p-value, and distance thresholds for P+T). The authors state that the benefit is in creating trans-ethnic score that perform well in all groups. However, Figure 3 is not showing necessarily the benefit of a trans-ethnic PRS, but potentially of a large sample size going into that trans-ethnic PRS. There are too many moving parts. It would benefit the conclusions to move the straight forward comparisons from the supplement to the main figures (Supp Figure 17A, for example).

8. This is not really necessary for any revisions or towards acceptance, but it would have been nice to see more exploration regarding the differential performance between cohorts with respect to prevalence of environmental risk factors, specifically since they have covariates available and the authors cite socioeconomic status as a major contributor to disparities.

Author Rebuttals to Second Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

I have no further comments.

The form of the BF now includes $\log(n)$ and a sufficient explanation is given.

Referee #4 (Remarks to the Author):

The authors have made substantial revisions to address the reviewers comments. They have focused on trans-ethnic analyses of lipids with respect to GWAS discovery, fine-mapping, and polygenic risk score development. Overall, the manuscript, specifically the results section, is very long and contains information that is not mentioned in the discussion or provided with an interpretation. Some of this can be cut out of the main text and left in the supplement. The manuscript could be tightened up some and distilled to its core points, but the necessity for this depends on the word count and formatting. I have listed specific concerns below.

Thank you. We have removed portions of the text and streamlined the remaining results as suggested by Reviewer 4 and the editor.

1. The authors conduct an analysis looking at the number of significant loci as a function of sample size, showing a linear trend. However, statistical power is not linear with respect to sample size, but exponential. Expectation should therefore fit an exponential, not linear, trend.

Response: We agree that statistical power for a single variant is not linear with sample size and is exponential when the effect size being discovered is held constant. However, the relationship between power and sample size can still appear quite linear when the effect sizes under investigation are very small (see for example https://www.power-analysis.com/effect_size.htm) or may be influenced by an uneven distribution of effect sizes.

Furthermore, we evaluate this empirically for lipids using a variety of smaller N subsets of the European data and plot the number of independent loci by sample size. Based on the data we have the fit of a curvilinear model is extremely poor. We are not the only group to have observed a linear relationship between sample size and number of loci, see for example Visscher et al, AJHG 2017 where they state “the typical linear relationship between sample size and number of loci has been observed” (in <https://doi.org/10.1016/j.ajhg.2017.06.005>, PMID: 28686856).

Assuming a non-infinitesimal polygenic model, we might eventually observe a decreasing number of discovered independent loci (using a 500kb distance definition of a locus) as we saturate the distance-based bins with associated loci.

We have simplified Supplemental Figure 7 to illustrate this better by focusing on total loci for LDL cholesterol and comparing subsets of the European meta-analysis to be in a similar sample size range as the non-European ancestries. The updated figure and legend now reads:

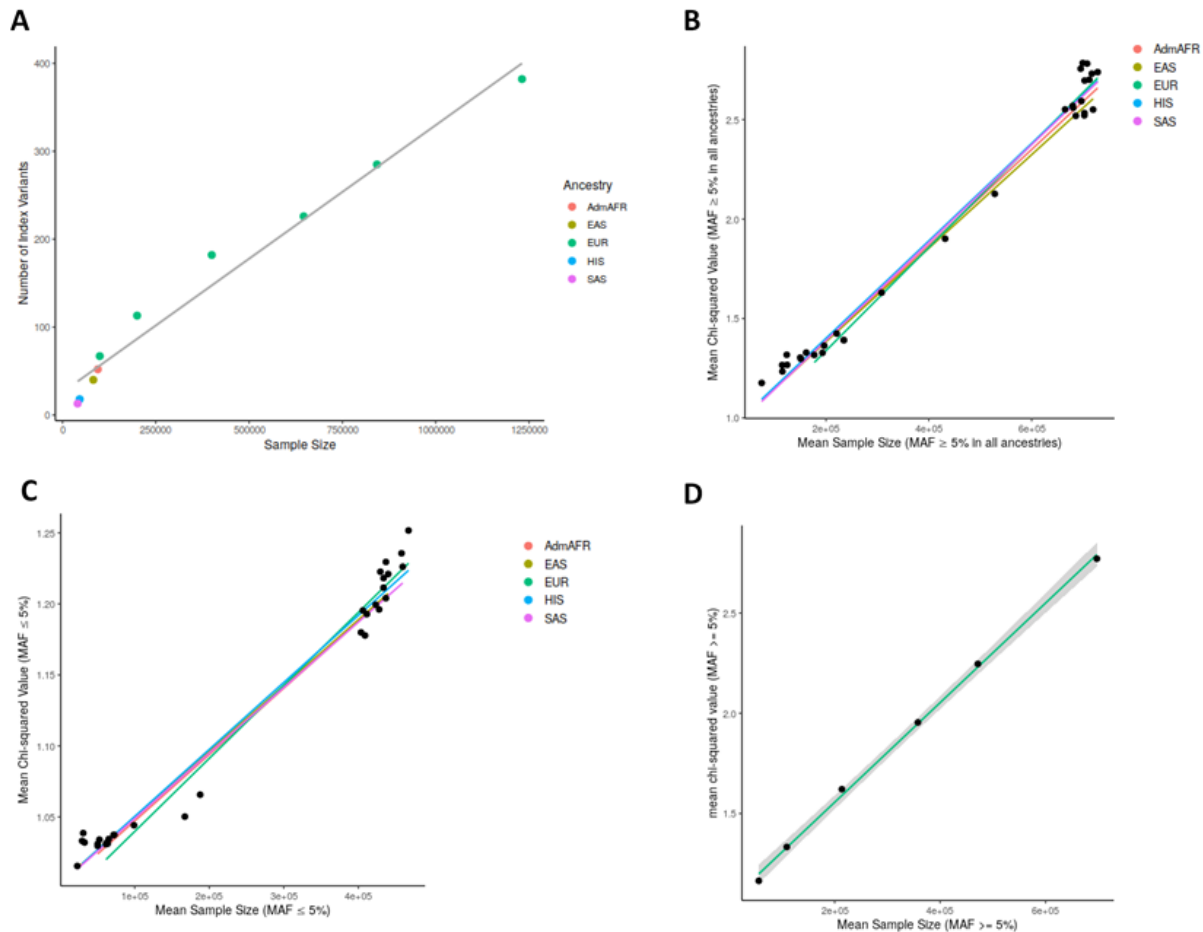
Supplementary Figure 7: Comparison of the number of index variants by sample size

A) Comparison of the number of index variants from meta-analysis of LDL in each ancestry group. A meta-analysis of five random subsets of European cohorts selected to reach sample sizes of approximately 100,000, 200,000, 400,000, 600,000, or 800,000 individuals is also shown.

B) Comparison of chi-squared values from meta-analysis for LDL of each possible combination of ancestry groups without genomic-control correction for variants with minor allele frequency (MAF) $\geq 5\%$. The colored lines indicate a linear regression model of all meta-analyses that include a given ancestry (eg. all analyses including European individuals).

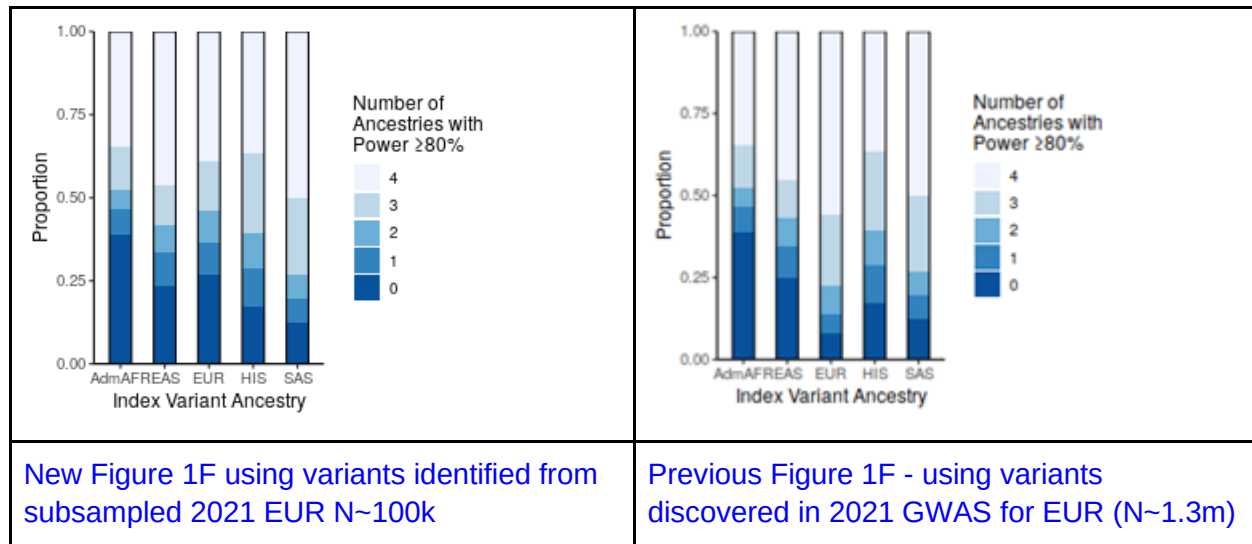
C) Comparison of chi-squared values from meta-analysis for LDL for variants with MAF $\leq 5\%$.

D) Comparison of chi-squared values for variants with MAF $\geq 5\%$ for LDL without genomic-control correction in a meta-analysis of all European cohorts combined as well as five subsets selected to reach sample sizes of approximately 100,000, 200,000, 400,000, 600,000, or 800,000 individuals.



2. Figure 1F shows the number of ancestries with power $>80\%$, keeping sample size constant. However, it seems that the index variants were selected on the stratified analyses, which have much different sample sizes. The authors have eliminated one aspect of difference (sample size for power calculations), but not the vast differences in discovery sample sizes that would confound

We agree that the variants discovered in the substantially larger European cohort may have different replicability in other ancestries due to differences in sample size and therefore power for discovery, to ancestry-specific effects, and to differences allele frequencies. To keep the sample sizes more consistent for the “discovered” variants which were then used to investigate power in other ancestries, we performed a meta-analysis for each trait with $\sim 100,000$ European individuals and used this to estimate power in other ancestries so that sample sizes are more comparable across groups. We have attached the new and ‘old’ figures below – you can see that after subsampling to 100k Europeans, the variants now appear to have a more similar level of ancestry-enrichment as for other studies with similar genetic diversity (namely EAS, HIS and SAS). This was a much-appreciated suggestion from the reviewer and substantially improves the comparison.



3. On pg 22 the authors assume a shared effect size across ancestries to estimate the proportion of variance explained. It is unclear why this assumption is made, considering that on the next page they show decreased correlation between ancestries with 7% reaching a P-value threshold for heterogeneity in effects. Why weren't the variants identified from the trans-ethnic analysis, but the population-specific weights used for this calculation?

We thank the reviewer for this suggestion. We have now repeated the estimates of the proportion of variance explained using population-specific weights. The section describing this result has been moved to the Supplement (per Editor request):

Original sentence in Main Text.

*Assuming a shared effect size across ancestries with varying allele frequencies, we find that the proportion of variance explained by the trans-ancestry index variants within each ancestry is 14%, 32%, 3%, 16%, and 20% lower on average in Admixed Africans, East Asians, Europeans, Hispanics, and South Asians, respectively (**Supplementary Table 7**).*

Revised sentence, now in Supplement

*Using population-specific effect size estimates and allele frequencies, we find that the proportion of variance explained by the trans-ancestry index variants within each ancestry on average is 28%, 11%, and 17% lower in Admixed Africans, East Asians, and South Asians and 1% and 6% higher in Europeans and Hispanics, respectively, relative to the trans-ancestry estimate (**Supplementary Table 19**). For variants successfully imputed into all ancestry groups, this corresponds to 8%, 9%, 11%, 11%, and 9% of the variance for Admixed African, East Asian, European, Hispanic, and South Asian individuals, respectively.*

4a. I found the last paragraph of the fine-mapping section confusing and unsupported with evidence (pg 26). Given that there is no functional follow-up to identify causal alleles, what evidence do the authors have that the indel within ANGPTL3 is the more likely causal variant compared to DOCK7?

In order to streamline the text and focus on the improvement from trans-ancestry analysis, we have removed the fine-mapping subsection that focused on individual ancestry results. This example is therefore no longer highlighted in the text. The credible sets in each ancestry are still included in **Supplementary Table 9**.

4b. Additionally, do the variants identified as having a high posterior probability in African populations have higher allele frequencies than in other populations? Or is the LD block broken up more?

We thank the reviewer for this suggestion. We assessed the LD patterns and frequency differences for index variants in association signals reaching genome-wide significance ($p\text{-value} < 5 \times 10^{-8}$) in both Admixed Africans and Europeans alone. We find that among association signals with improved fine-mapping, 92% have fewer variants in high LD ($r^2 > 0.8$) in Africans (based on 1000 Genomes variants) compared to Europeans, while only 56% of the index variants at these signals were more common in Admixed Africans in our dataset. Therefore, the improved fine-mapping appears to be driven by LD differences more so than frequency differences. This analysis is now included in the **Supplementary Note** and in **Supplementary Table 20**.

(New text in the Supplementary Note)

In order to quantify the improvement in fine-mapping through trans-ancestry meta-analysis, we grouped the 2,286 index variants into 1,486 independent association signals based on an LD r^2 threshold of 0.7 between index variants. This was done to avoid double-counting overlapping association signals. Considering all independent association signals under the assumption of a single, shared causal variant, we found a median 40% reduction in credible set size for regions with improved fine-mapping in the trans-ancestry meta-analysis. We next aimed to determine whether differences in linkage disequilibrium patterns or allele frequency differences were driving this improvement. Starting with the independent association signals, we selected for further analysis the 151 signals that reached a significance threshold of $p\text{-value} < 5 \times 10^{-8}$ in both the Admixed African and European meta-analyses alone. For each of these regions, we manually inspected LocusZoom plots from the Admixed African and European ancestry-specific meta-analyses and from trans-ancestry meta-analysis to exclude any loci with apparent secondary signals within the region. Of the 69 association signals selected for further analysis, 36 (52%) had the smallest 99% credible set from the trans-ancestry meta-analysis, 6 (9%) from the Admixed African ancestry meta-analysis, 3 (4%) from the European ancestry meta-analysis, and 24 (35%) had equivalent fine-mapping in two or more analyses. Among the 36 signals with improved fine-mapping, we observed a median 50% reduction in credible set size.

We next focused our comparisons on the 36 association signals with improved fine-mapping in the trans-ancestry meta-analysis. The trans-ancestry index variants at these signals were more common in Admixed Africans than Europeans 56% of the time (20/36), with a median 1.3-fold difference. We then identified all variants in 1000 Genomes that were in strong linkage disequilibrium ($r^2 > 0.8$) with the trans-ancestry index variant in Africans or Europeans. Nearly all loci (33/36, 92%) with improved fine-mapping had fewer variants in high LD among Africans compared to Europeans. For example, in 1000 Genomes Africans there was a median of 6 variants having $r^2 > 0.8$ with the trans-ancestry index variant compared to a median of 40 variants having $r^2 > 0.8$ in Europeans. Therefore, the improved fine-mapping observed in the trans-ancestry meta-analysis appears to be primarily due to the smaller number of variants in high LD with the lead index variant in Africans relative to Europeans rather than differences in allele frequency between populations.

4c. I would also be hesitant to say that these are ancestry-specific to Africa, given the authors only compare Africa to Europe. Therefore it may be a non-European-specific trend.

[redacted]

We have removed this sentence from the revised text and instead highlight the reduction of credible set size from the overall trans-ancestry meta-analysis rather than discussing variants observed in Admixed Africans specifically.

5. The authors examine the relationship between their trans-ancestry score and the principal components in 1000Genomes. It is unclear why they did not do this in their data that includes outcome data to investigate if this relationship is relevant to prediction or a possible source of bias.

We thank the reviewer for this suggestion. The comparison between the trans-ancestry score and principal components was suggested during a previous round of revisions as correlation between polygenic scores and principal components has been observed for other traits. We have now compared the prediction of LDL-C within European and African American individuals in the Michigan Genomics Initiative using the Admixed African, European, or trans-ancestry scores either alone or in combination with covariates and varying numbers of principal components included in the model. These results are given in **Supplementary Tables 22 and 23** and in the **Supplementary Note**. We find that the prediction as measured by adjusted R^2 is similar when varying numbers of PCs are included and the trans-ancestry score remains the best predictor of LDL-C in both European and African American individuals. The revised text is included below:

(Text in the Supplementary Note)

*Previous studies have suggested that population stratification may influence the predictive ability of polygenic scores across diverse populations²⁶. We tested for correlation between the trans-ancestry polygenic score and principal components of ancestry (PCs) in 1000 Genomes individuals. Significant correlation was observed between the trans-ancestry polygenic score and PCs 1 and 3 only (p -value < 0.0025; 0.05/20 tested PCs; **Supplementary Table 21**). Within the MGI cohort, we found that median LDL-C values were not significantly correlated with any of the first twenty principal components (p -value > 0.0025) and prediction of LDL-C as measured by adjusted R^2 was similar when either PCs 1-4, 1-10, or 1-20 were included as covariates in the model with the ancestry-specific or trans-ancestry polygenic scores (**Supplementary Tables 22 and 23**). We note that normalization of LDL polygenic scores should be performed within each ancestry.*

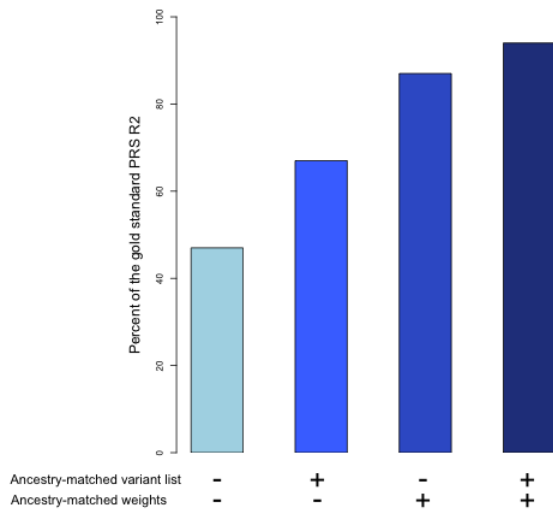
6. I found the second to last paragraph of the PRS section confusing. The authors offer solutions of how to improve PRS performance when starting with EUR-derived scores. However, their suggestions seem to be to either re-select variants based on parameters in your target population and/or use weights from your target population. This seems like they are saying to basically create a new score in this new population, but I could be misunderstanding this.

In the absence of an adequately powered ancestry matched GWAS, researchers interested in polygenic prediction must figure out the optimal way to adapt existing data for prediction in the target population. It is currently unknown whether: (i) ancestry-matched lists of variants or (ii) ancestry-specific effect sizes are most crucial to successful polygenic prediction in understudied ancestries.

We have reduced the text for clarity (now in the Supplementary Note and Supplementary Figure 17). Starting from a completely ancestry mismatched score that only captures 47% of the PRS performance of an ancestry-matched score, we describe the impact of first updating the variant list based on an ancestry-specific GWAS (i.e. a small N ancestry-specific GWAS exists or ancestry-specific LD and frequency were leveraged to develop a more informative variant list) yet the weights remain mismatched by ancestry. Next, we performed the opposite -- updated the weights with ancestry-matched summary statistics but the variant list remained the same (i.e. an array of PRS variants was genotyped in another ancestry). We report how much of the optimal PRS performance (ancestry-matched variant lists and weights) was gained (71% and 87%, respectively).

(New) Supplementary Figure 17: *Improvement in PRS performance in African Americans when starting with ancestry-mismatched European scores by updating weights or updating variant lists.*

By comparison to the gold-standard score of the trans-ancestry-derived PRS (Adjusted $R^2 = 0.12$), a European-derived score captures only 47% of the variance explained by the trans-ancestry PRS. When LD and association information from the target population is used to optimize the list of variants for inclusion in the PRS, but with ancestry-mismatched weights from Europeans, the variance explained reaches 71% of the gold standard. If the PRS variant list selected in Europeans were genotyped in the target population, and PRS weights were updated using a GWAS from the target population, the variance explained reached 87% of the gold standard. Finally, deriving both the marker list and weights from the target population (single-ancestry) explained 94% of the variance relative to the gold-standard trans-ancestry PRS.



Text from Supplementary Note:

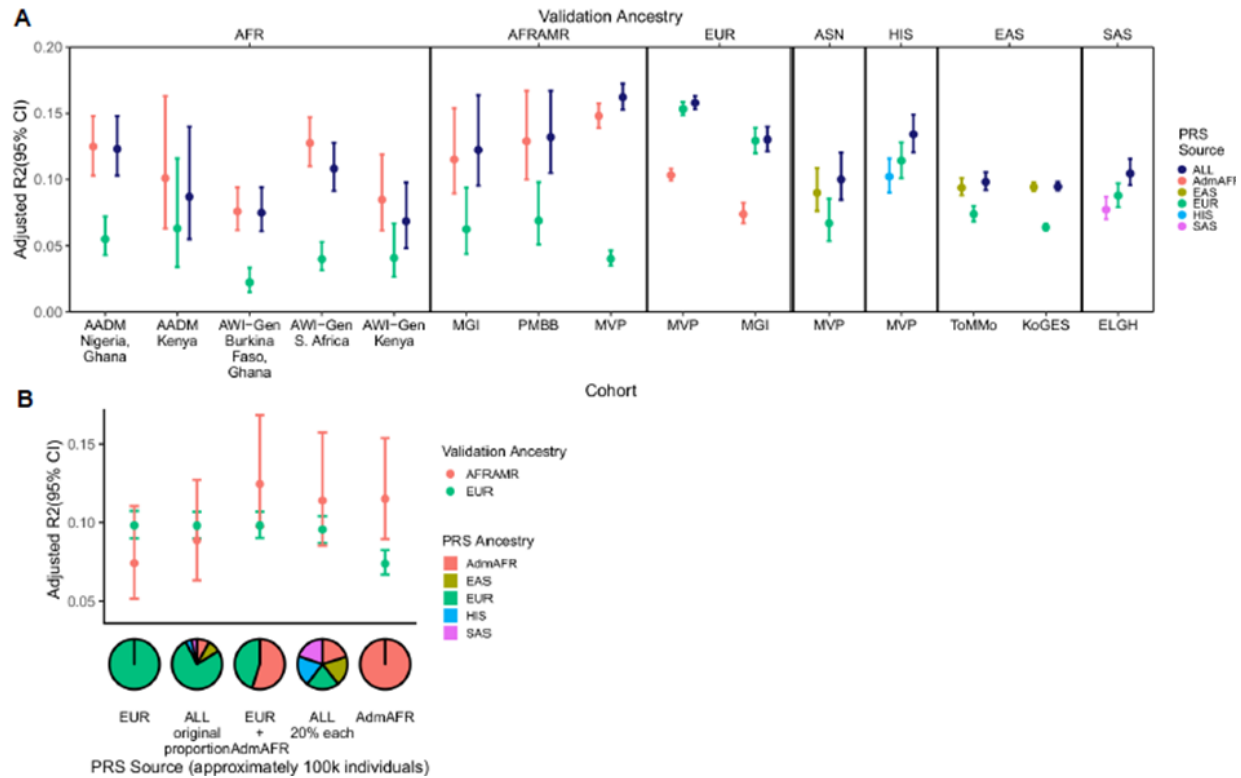
Finally, we aimed to investigate the role of variant selection and weights in polygenic score performance. Poor performance of ancestry mismatched scores could be caused by either missing ancestry-specific variants in the score or by differing LD with the underlying causal variant between ancestry groups leading to imperfect variant weights. Starting with ancestry mismatched pruning and thresholding scores, we attempted to ‘correct’ the ancestry mismatch by first applying ancestry-matched weights. This helps with future study design questions – e.g. would a single pre-defined set of variants on an array be useful for all ancestries if we applied updated ancestry-specific weights? We used the predictive ability of the pruning and thresholding trans-ancestry score as the ‘gold standard’ because it achieved the highest R^2 for any polygenic score. In admixed African individuals, we recovered 87% of the gold standard polygenic score when we used the European variant list with admixed African weights compared to just 47% when using the European variant list with weights from Europeans (Supplementary Figures 15 and 17, Supplementary Table 17).

We then examined the role of optimizing variants selected for polygenic scores. We found that European ancestry GWAS-derived score had improved prediction in

individuals with admixed African ancestry when the variant selection parameters (such as p-value thresholds for pruning and threshold) were selected based on optimizing the score in admixed African ancestry rather than in Europeans. Using parameters optimized from only European individuals led to prediction in admixed African individuals that was just 47% of the gold standard while using parameters optimized in admixed African individuals (a more stringent p-value threshold of 5×10^{-10}) resulted in prediction that was 67% of the gold standard (Supplementary Table 17), even with European weights. Finally, using the AFR ancestry-matched weights and ancestry-matched variant list from a single-ancestry AFR GWAS resulted in 94% of the gold standard (trans-ancestry) polygenic score performance. Taken together, our findings suggest that polygenic scores derived from ancestry-mismatched GWAS may be improved by substituting ancestry-specific weights for the selected variants when ancestry-matched GWAS of sufficient sample sizes are not available, and/or by optimizing the variant selection in ancestry-matched individuals.

7. While the authors have added supplemental materials that address some of this concern, the message is still a bit muddled when it comes to PRS performance, given the different methods (P+T, PRS-CS) and different parameters (r^2 , p-value, and distance thresholds for P+T). The authors state that the benefit is in creating trans-ethnic score that perform well in all groups. However, Figure 3 is not showing necessarily the benefit of a trans-ethnic PRS, but potentially of a large sample size going into that trans-ethnic PRS. There are too many moving parts. It would benefit the conclusions to move the straight forward comparisons from the supplement to the main figures (Supp Figure 17A, for example).

We agree with the reviewer that the PRS developed from matched sample sizes is necessary to demonstrate the utility of a trans-ancestry score and have therefore moved (old numbering) Supplemental Figure 17A to main figure 3B. We believe both analyses are valid and useful - the first in which the PRS score from the most well-powered GWAS (as is reflective of many other European-biased GWAS studies) is optimized for performance and tested in many ancestries, and the second one proposed by R4 in which the optimization parameters are held more constant (including GWAS N).



8. This is not really necessary for any revisions or towards acceptance, but it would have been nice to see more exploration regarding the differential performance between cohorts with respect to prevalence of environmental risk factors, specifically since they have covariates available and the authors cite socioeconomic status as a major contributor to disparities.

We agree this would be very interesting, but unfortunately it would also be difficult to answer in a timely manner or one consistent with word constraints. The covariates examined at baseline for the PRS-testing cohorts were age, sex, and principal components of ancestry. The analyses in this paper are from summary statistics from ~170 primary cohorts where each cohort performs analyses for new scientific questions, followed by centralized QC, meta-analysis, interpretation and write-up of results. We do not have immediate access to individual-level data e.g. covariates from more than a few of the primary studies. Even if we did have access to covariate data for more of the studies, a complex, time consuming harmonization effort for environmental variables across cohorts would need to precede the analyses requested.

Reviewer Reports on the Third Revision:

Referee #4 (Remarks to the Author):

The authors have further streamlined their work and greatly improved the manuscript to focus on their main points, which contribute greatly to the field. In my opinion this also increases the impact of the manuscript with a very clear message. They have addressed all of my comments and I have no further concerns regarding the manuscript.

The following points are very very minor and are only for clarifying specific points if possible given word counts and relevancy.

- Page 21 lines 868-871 states that the numbers are variance explained per standard deviation in polygenic score, but the numbers themselves look like effect sizes per SD. I may be misreading this, but just in case they are effect sizes it would be nice to also have the CI to further conclude that they are similar between ancestry groups.
- I recognize that this can get wordy, but there are some inconsistencies in population descriptions (Europeans, European-ancestry, EUR).
- If you have the word count bandwidth, it would be helpful to expand what correlations between traits in UKBB and MVP mean (just some context for the reader). This is page 19 line 808.
- It is unclear why African and Hispanic ancestry-specific analyses have most loci per genotyped individual due to recent admixture, instead of because of having shared African ancestry (and therefore a site frequency spectrum shifted compared to other groups) on page 18 lines 769-771.

Author Rebuttals to Third Revision:

Referee #4 (Remarks to the Author):

The authors have further streamlined their work and greatly improved the manuscript to focus on their main points, which contribute greatly to the field. In my opinion this also increases the impact of the manuscript with a very clear message. They have addressed all of my comments and I have no further concerns regarding the manuscript.

Thank you for your careful review and your previous comments which have undoubtedly improved the manuscript.

The following points are very very minor and are only for clarifying specific points if possible given word counts and relevancy.

- Page 21 lines 868-871 states that the numbers are variance explained per standard deviation in polygenic score, but the numbers themselves look like effect sizes per SD. I may be misreading this, but just in case they are effect sizes it would be nice to also have the CI to further conclude that they are similar between ancestry groups.

We apologize that the previous wording was unclear. The values given are effect sizes. We have edited this sentence to clarify that the values are the increase in LDL-C per standard deviation and have added the standard errors. This sentence now reads:

The increase in LDL-C per each standard deviation increase in the polygenic score was also similar between ancestry groups in MVP: 13.2±0.22 mg/dL for African American, 8.9±0.47 mg/dL for Asian (EAS/SAS), 10.5±0.10 mg/dL for European, and 10.6±0.32 mg/dL for Hispanic ancestry individuals

- I recognize that this can get wordy, but there are some inconsistencies in population descriptions (Europeans, European-ancestry, EUR).

We have edited the population descriptions throughout the text to use European-ancestry rather than European to reflect the fact that our European-ancestry participants were recruited both within Europe and elsewhere (eg. United States). We have also edited the text to minimize the use of abbreviations for ancestry groups.

- If you have the word count bandwidth, it would be helpful to expand what correlations between traits in UKBB and MVP mean (just some context for the reader). This is page 19 line 808.

We thank the reviewer for this suggestion. We have added an explanation to this section to note that these correlations suggest moderately high sharing of effect sizes between ancestry groups. This portion of the text now reads:

*Genetic correlation between Admixed African and European ancestries for HDL-C ($r=0.84$) was not significantly different from 1 in the UK Biobank (possibly due to relatively small numbers of African ancestry individuals), while correlations for the other traits ranged from 0.52-0.60 in UK Biobank and 0.47-0.69 in MVP (**Supplementary Table 8**). These results indicate moderately high correlation in lipid effect sizes across ancestry groups when considering all genome-wide variants.*

- It is unclear why African and Hispanic ancestry-specific analyses have most loci per genotyped individual due to recent admixture, instead of because of having shared African ancestry (and therefore a site frequency spectrum shifted compared to other groups) on page 18 lines 769-771.

We have edited this portion of the text to note that both African ancestry and Indigenous ancestry may influence the number of identified loci. This sentence now reads:

*Next, we compared the number of loci identified per 100,000 participants in each ancestry group and the combined dataset (**Figure 1E**). African and Hispanic ancestry-specific analyses identified the most loci per genotyped individual, perhaps due to shared African ancestry and/or the presence of substantial Indigenous ancestry among Hispanic ancestry individuals; European and trans-ancestry analyses identified slightly fewer loci per 100,000 individuals, likely reflecting a slight reduction in the benefit from new samples added to very large sample sizes (>1m).*