Peer Review Information

Journal: Nature Genetics
Manuscript Title: Leveraging fine-mapping and multi-population training data to improve cross-population polygenic risk scores
Corresponding author name(s): Dr Omer Weissbrod, Alkes L. Price

Reviewer Comments & Decisions:

Decision Letter, initial version:

17th Mar 2021

Dear Dr Weissbrod,

Your Article, "Leveraging fine-mapping and non-European training data to improve trans-ethnic polygenic risk scores" has now been seen by 2 referees. You will see from their comments below that while they find your work of interest, some important points are raised. We are interested in the possibility of publishing your study in Nature Genetics, but would like to consider your response to these concerns in the form of a revised manuscript before we make a final decision on publication.

To guide the scope of the revisions, the editors discuss the referee reports in detail within the team, with a view to identifying key priorities that should be addressed in revision. As you will see from these comments, referees are generally positive about the PolyPred and PolyPred+ methods and the utility for cross-ancestry PRS optimization. Both referees have identified aspects of the analyses and the methodological details that need to be improved or clarified. We therefore invite you to revise your manuscript taking into account all reviewer comments. Please highlight all changes in the manuscript text file. At this stage we will need you to upload a copy of the manuscript in MS Word .docx or similar editable format.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

When revising your manuscript:

*1) Include a “Response to referees” document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the referees along with the revised manuscript.

*2) If you have not done so already please begin to revise your manuscript so that it conforms to our Article format instructions, available
<a href="http://www.nature.com/ng/authors/article_types/index.html">here</a>.
Refer also to any guidelines provided in this letter.

*3) Include a revised version of any required Reporting Summary:
https://www.nature.com/documents/nr-reporting-summary.pdf
It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer review.
A revised checklist is essential for re-review of the paper.

Please be aware of our <a href="https://www.nature.com/nature-research/editorial-policies/image-integrity">guidelines on digital image standards.</a>

Please use the link below to submit your revised manuscript and related files:

[REDACTED]

<strong>Note:</strong> This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We hope to receive your revised manuscript within four to eight weeks. If you cannot send it within this time, please let us know.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Nature Genetics is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as ‘corresponding author’ on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on ‘Modify my Springer Nature account’. For more information please visit please visit <a href="http://www.springernature.com/orcid">www.springernature.com/orcid</a>.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Sincerely,

Wei Li, PhD
Senior Editor
Nature Genetics
One New York Plaza, 47th Fl.
New York, NY 10004, USA
www.nature.com/ng

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
In “Leveraging fine-mapping and non-European training data to improve trans-ethnic polygenic risk scores" by Weissbrod et al., the authors propose new methods, PolyPred and PolyPred+, for polygenic risk scores in trans-ancestry populations. The authors compare these new methods to multiple state-of-the-art methods, demonstrate the utility of these new methods in simulations and across many traits in multiple datasets, and provide extensive "secondary" analyses to explore a variety of scenarios. The manuscript is well written, and the study is comprehensive and thorough. The authors clearly demonstrate the power of these new methods and the importance of these approaches for using polygenic risk scores in non-European populations.

In the first sentence of the abstract, as well as articulated later in the text, the authors make the point that PRS based on European training data suffers reduced accuracy in non-European populations. Do we KNOW this to be true? What is the evidence? Is it really because the sample size in the non-European populations is so much smaller and this is reducing our power? I am not sure that I have seen properly powered datasets from non-European populations being used to generate the GWAS summary statistics from which to derive the weights for the PRS. While these new methods do a great job of improving upon PRS for more diverse datasets, I worry that the field continues to propagate this statement without evidence. Is this truly an issue of the underlying LD structure and heritability of the different ancestry groups – or is this an issue of power?

Similarly, in the second sentence of the abstract as well as the 3rd sentence of the introduction, the authors discus the drivers of the disparity in PRS methods, but fail to mention sample size. Based on the analyses in this paper alone, I would say that sample size could be driving the differential accuracy of PRS. Do any of the papers they cite (refs 11-27) explore this issue of sample size of the training data?

In the 2763 overlapping genomic regions, are these equally spaced starting at chromosome 1, position 0 or are these regions defined in some other more sophisticated manner?

In the section describing the simulations, the authors indicate what summary LD was used for each method. Why did you change the summary LD for all of the methods? How would the results change if you used the same summary LD across all of the methods? What impact does the summary LD have on the accuracy of the methods?

Also in the simulation the authors describe “Using UK10K was almost as accurate as using in-sample LD, but using 1000 Genomes Europeans was hugely inaccurate, leading to prediction accuracy even lower than that of P+T (Supplementary Table 1), confirming the importance of using a large (population-matched) LD reference panel to compute PRS41 (on the other hand, using UK10K LD is not recommended when using PolyFun for fine-mapping29 due to concerns about false-positive fine-mapped SNPs, which are not a primary concern when computing PRS).” Did you also try a downsampled set of UK10K of the same size as the 1000 Genomes Europeans (N=-489) to confirm whether it is primarily sample size or the underlying LD patterns that are driving the results? I expect the 1000 Genomes Europeans are of more diverse ancestry that the 3500 UK10K. So, is the sample size issue creating more smaller LD blocks than the UK10K? Would it look more similar if both are less than 500 people to estimate the LD?

The compute time differences are quite large (2.8 minutes versus 668 minutes). What is causing the 300-times longer compute? How much of an impact will this have on the ability to use PolyPred in real-world scenarios?

At the end of the results section, the authors say “We emphasize that efforts to assess the benefit of incorporating non-European training data should analyze non-European training data from a cohort that is distinct from the target cohort, otherwise results may be inflated due to cohort effects.” This is perhaps one of the most critically important points of the paper. Does this mean that the use of very large, trans-ancestry meta-analyses to estimate the weights in the training data would be optimal as long as the target cohort is not included? Did you evaluate any of the traits using this approach where a meta-analysis of many studies and/or ancestry groups is the basis of the training data?

In the discussion, the authors state “Finally, PRS may implicitly capture GxE interactions, which may not be transferable across cohorts or ancestries27,75.” This is an interesting point. But why this would be the case is not clear based on any of the analyses provided here. Can this be explained further? How are the GxE effects being captured by PRS and more importantly, how can we determine whether they are transferable across ancestry?

In the methods, where do you get the 187 functional annotations for Polyfun-pred?

Throughout the paper, the authors use the term “ethnicity”, including in the figures. Is “ethnicity” the correct word? It seems as though the focus is really on genetic ancestry, which is different from ethnicity. As we try to move as a field toward removing racism and bias, it seems like being more precise with our words, such as the use of the word “ancestry” when we are truly using genetics to define groups, seems more appropriate. If I am misunderstanding the way that groups are defined herein, and ethnicity is more accurate, please keep ethnicity but consider adding an explanation in the paper as to why that word is being used.

Reviewer #2:
Remarks to the Author:
This manuscript from Weissbrod, Price and colleagues present an extension of a set of popular stats gen/fine mapping technologies, combining them to maximise prediction in non European populations. This work is topical and relevant given the excitement around cross-ancestry PRS optimisation. The results do seem convincing, and on that basis I am rather positive about the work, but I struggled to go through the manuscript and I do think that it could be improved substantially.

First of all, this manuscript is quite technical and it took me some time to understand its flow. An additional complication is that not all methods are usable without individual level data, which further complicates the choice of methodology. These points are discussed in various sections of the manuscript, but a key comment would be whether the authors would consider a flowchart of some sort that summarises how the various methods fit together, with or without the availability of individual level data. I am aware that saying that a manuscript is hard to follow is a typically unhelpful reviewer's comment, but I think it's fair to say that the ratio text:display items is high, and additional help to understand how it all fits would help. More sub-headers would perhaps also be helpful, as some sections are very lengthy and cover a lot of ground.

Among these complications, the requirement for individual level data is in fact quite critical, as one may argue that the comparisons are somewhat academic given that the use of summary statistics typically yields much larger training sets. It does not mean that the insights from this paper aren't important, but I would like this individual level data requirement to be made clearer (for example in the introduction). It is obviously another layer of complication but I would like to see what Polypred-S is doing in the main display items (figures 2-4).

My last major point relates to the general methodology of linearly combining different PRS. Fundamentally, BOLT-LMM and Polypred-fun are capturing much of the same signal, so there is some double counting and some single counting of signals, which has to be sub-optimum. I would argue that some form of signal subtraction would make more sense. I wonder if the authors could comment on that point. Dealing with this issue properly is probably beyond the scope of this manuscript, but I think it would be helpful to consider what could be done.

Lastly, on a more minor point, the authors speak well of the SBayesR results outperforming PRS-CS and LDPred. This puzzles me, as I must say I never managed to get SBayesR to run properly. My struggles are echoed by the recent bioRxiv manuscript from Pain, Lewis and colleagues which I am sure the authors are aware of (https://doi.org/10.1101/2020.07.28.224782). I'd be keen to understand why such stark differences are being observed.

Author Rebuttal to Initial comments

Reviewer #1: (numbers added to reviewer comments)

In “Leveraging fine-mapping and non-European training data to improve trans-ethnic polygenic risk scores" by Weissbrod et al., the authors propose new methods, PolyPred and PolyPred+, for polygenic risk scores in trans-ancestry populations. The authors compare these new methods to multiple state-of-the-art methods, demonstrate the utility of these new methods in simulations and across many traits in multiple datasets, and provide extensive "secondary" analyses to explore a variety of scenarios. The manuscript is well written, and the study is comprehensive and thorough. The authors clearly demonstrate the power of these new methods and the importance of these approaches for using polygenic risk scores in non-European populations.

We thank the reviewer for the accurate summary, and for suggesting that the power and importance of PolyPred and PolyPred+ for PRS in non-European populations has been clearly demonstrated.

1. In the first sentence of the abstract, as well as articulated later in the text, the authors make the point that PRS based on European training data suffers reduced accuracy in non-European

populations. Do we KNOW this to be true? What is the evidence? Is it really because the sample size in the non-European populations is so much smaller and this is reducing our power? I am not sure that I have seen properly powered datasets from non-European populations being used to generate the GWAS summary statistics from which to derive the weights for the PRS. While these new methods do a great job of improving upon PRS for more diverse datasets, I worry that the field continues to propagate this statement without evidence. Is this truly an issue of the underlying LD structure and heritability of the different ancestry groups – or is this an issue of power?

The reviewer has raised 2 related questions: (a) is there compelling evidence that PRS based on European training data suffer reduced accuracy in non-European populations?; and (b) is the poor performance of PRS in non-European populations due to small non-European sample size (which reduces power) or to LD and heritability differences? We address each question in turn.

(a) Is there compelling evidence that PRS based on European training data suffer reduced accuracy in non-European populations?

We believe that there is compelling evidence that PRS based on European training data suffer reduced accuracy in non-European populations. For example, Martin et al. 2019 Nat Genet (ref. 13) reported that prediction in African individuals based on European-derived summary statistics suffered from a 4.5x loss of accuracy vs Europeans. Similar findings were reported in Marquez-Luna et al. 2017 Genet Epidemiol (ref. 7), Duncan et al. 2019 Nat Commun (ref. 11), Wang et al. 2020 Nat Commun (ref. 14), Amariuta et al. 2020 Nat Genet (ref. 15), Marnetto et al. 2020 Nat Commun (ref. 16), and Chen et al. 2020 Cell (ref. 18). However, we recognize that it is our responsibility to provide a clear exposition. Accordingly, we have updated the Introduction section (p.2) to expand the description of published work showing that PRS based on European training data suffer reduced accuracy in non-European populations.

We note that the above response does not pertain to PRS based on non-European training data (but see (b) below).

(b) Is the poor performance of PRS in non-European populations due to small non-European sample size (which reduces power) or to LD and heritability differences?

We would like to emphasize the distinction between using *European training data* vs. *non-European training data* to compute PRS in non-European populations.

If *European training data* is used, the poor performance is due to LD and heritability differences. In particular, the small non-European target sample size is not relevant, as only the training sample size impacts power. We have verified this in a new experiment in which we downsampled the UK Biobank (non-British) European target sample to match the size of the UK Biobank African target sample; as expected, PRS accuracy remained much larger (+276% larger) in the (non-British) European target sample vs. the African target sample, despite the matched target sample sizes. Results are reported in Supplementary Table 4, cited in the *Analysis of 4 UK Biobank populations using UK Biobank British training data* subsection of the Results section (p.9). We further note that although small target sample sizes do not impact PRS accuracy, they can lead to noisy estimates of PRS accuracy. However, the differences that we observe in PRS accuracies between European vs. non-European target cohorts are statistically significant, and we now provide p-values for these differences in Supplementary Table 4 and Supplementary Table 6, cited in the *Analysis of 4 UK Biobank populations using UK Biobank British training data* subsection of the Results section (p.7).

If *non-European training data* is used, the poor performance is due to small non-European training sample size (which reduces accuracy). Indeed, the dependence of PRS accuracy on training sample size has previously been demonstrated in Chatterjee *et al.* 2016 Nat Rev Genet (ref. 1), Gurdasani *et al.* 2019 Nat Rev Genet (ref. 12), Martin *et al.* 2019 Nat Genet (ref. 13), Amariuta *et al.* 2020 Nat Genet (ref. 15), Mills *et al.* 2020 Nat Genet (ref. 21), Vilhjalmsson *et al.* 2015 Am J Hum Genet (ref. 33).

Again, we recognize that it is our responsibility to provide a clear exposition. We have updated the Abstract (p.1) and Introduction section (p.2) to clarify these points.

2. Similarly, in the second sentence of the abstract as well as the 3rd sentence of the introduction, the authors discuss the drivers of the disparity in PRS methods, but fail to mention sample size. Based on the analyses in this paper alone, I would say that sample size could be driving the differential accuracy of PRS. Do any of the papers they cite (refs 11-27) explore this issue of sample size of the training data?

(also see part (b) of response to Reviewer #1 Comment 1)

Again, we would like to emphasize the distinction between using *European training data* vs. *non-European training data* to compute PRS in non-European populations.

If *European training data* is used, the small non-European test sample size is not relevant, as only the training sample size impacts power.

If *non-European training data* is used, the poor performance is due to small non-European training sample size (which reduces accuracy). Indeed, the dependence of PRS accuracy on training sample size has previously been demonstrated in Chatterjee *et al.* 2016 Nat Rev Genet (ref. 1), Gurdasani *et al.* 2019 Nat Rev Genet (ref. 12), Martin *et al.* 2019 Nat Genet (ref. 13), Amariuta *et al.* 2020 Nat Genet (ref. 15), Mills *et al.* 2020 Nat Genet (ref. 21), Vilhjalmsson *et al.* 2015 Am J Hum Genet (ref. 33).

As noted above, we have updated the Abstract (p.1) and Introduction section (p.2) to clarify these points.

3. In the 2763 overlapping genomic regions, are these equally spaced starting at chromosome 1, position 0 or are these regions defined in some other more sophisticated manner?

The 2,763 overlapping genomic regions are equally spaced starting at chromosome 1, position 0 (the definition of these genomic regions was first described in Weissbrod et al. 2020 Nat Genet (ref. 35)). We have updated the *PolyPred and its summary statistic-based analogues* subsection of the Methods section (p.17) to clarify this point.

4. In the section describing the simulations, the authors indicate what summary LD was used for each method. Why did you change the summary LD for all of the methods? How would the results change if you used the same summary LD across all of the methods? What impact does the summary LD have on the accuracy of the methods?

The reviewer has raised 3 related questions: (a) why did we use different summary LD data for different methods?; (b) what would happen if we use the same summary LD data for all methods?; and (c) what is the impact of summary LD data on PRS accuracy? We address each question in turn.

(a) Why did we use different summary LD data for different methods?

The reviewer is correct that we used different summary LD data for different methods. Specifically, PolyFun-pred, SBayesR, and PRS-CS, three of the methods included in our primary comparisons (we have added PRS-CS as a main method throughout our revised manuscript; see response to Reviewer #2 Comment 5), make use of summary LD in European training data. PolyFun-pred (the first component of PolyPred and PolyPred+) uses summary LD for all 18 million SNPs with MAF≥0.1% in the European training data (following the recommendations of Weissbrod et al. 2020 Nat Genet (ref. 35), the paper that introduced the PolyFun method). SBayesR uses summary LD for 1.2 million HapMap 3 SNPs (following the recommendations and publicly available LD matrices of Lloyd-Jones et al. 2019 Nat Commun (ref. 38), the paper that introduced the SBayesR method).

PRS-CS also uses summary LD for 1.2 million HapMap 3 SNPs (following the recommendations and publicly available LD matrices of Ge et al. 2019 Nat Commun (ref. 39), the paper that introduced the PRS-CS method).

PolyFun-pred, SBayesR, and PRS-CS use different algorithms to impose sparsity on LD matrices, and different file formats to store them. Thus, to run SBayesR or PRS-CS using summary LD from the same 18 million SNPs used by PolyFun-pred would require rerunning the summary LD computation pipeline of the respective methods (SBayesR or PRS-CS) from scratch. However, we believe that this would be computationally infeasible, based on the information provided in Lloyd-Jones *et al*. 2019 Nat Commun (ref. 38). Similarly, we believe that it would be computationally infeasible to run PRS-CS using 18 million SNPs, based on the information provided in Ge *et al.* 2019 Nat Commun (ref. 39). It is also not technically possible in the case of PRS-CS, because the authors of PRS-CS have not released software to compute LD matrices. We have updated the *Analysis of 4 UK Biobank populations using UK Biobank British training data* subsection of the Results section (p.8) and the Discussion section (p.15) to clarify these points.

(b) What would happen if we use the same summary LD data for all methods?

As noted in (a) above, we believe that running SBayesR or PRS-CS using summary LD from the same 18 million SNPs used by PolyFun-pred would be computationally infeasible, based on the information provided in Lloyd-Jones *et al.* 2019 Nat Commun (ref. 38) and in Ge *et al.* 2019 Nat Commun (ref. 39), and we have updated the *Analysis of 4 UK Biobank populations using UK Biobank British training data* subsection of the Results section (p.8) and the Discussion section (p.15) to clarify this point.

However, we have run SBayesR using summary LD from a SNP set larger than 1.2 million SNPs. In detail, in addition to the summary LD for 1.2 million HapMap 3 SNPs publicly released by Lloyd-Jones et al. 2019 Nat Commun (ref. 38), those authors have also publicly released summary LD for 2.8 million SNPs (pruned UK Biobank SNPs). We have performed secondary analyses using the summary LD for 2.8 million SNPs (SBayesR-2.8M) instead of summary LD for 1.2 million HapMap 3 SNPs (SBayesR). We determined that SBayesR-2.8M was less accurate than SBayesR (significantly so for Africans) (Supplementary Table 5, cited in the *Analysis of 4 UK Biobank populations using UK Biobank British training data* subsection of the Results section (p.8)). (These results do not contradict the findings of Lloyd-Jones et al. 2019 Nat Commun (ref. 38), who analyzed only 2 of 12 real traits using summary LD from 2.8 million SNPs, and reported that using summary LD from 2.8 million SNPs instead of summary LD from 1.2 million HapMap 3 SNPs improved prediction accuracy for 1 of 2 traits). Thus, there is little reason to believe that further expanding the SNP set used to compute summary LD would improve the performance of SBayesR. We note that we use SBayesR (instead of SBayesR-2.8M) in all primary comparisons, which is a conservative choice since SBayesR outperforms SBayesR-2.8M. We have updated

the *Analysis of 4 UK Biobank populations using UK Biobank British training data* subsection of the Results section (p.8) to clarify this point.

As an alternative, we performed simulations in which we trained PolyFun-pred using only 1.2 million HapMap3 SNPs (instead of 18 million MAF>0.1% SNPs) and obtained substantially and significantly reduced accuracy (-11% relative-R^2 in non-British Europeans vs. using all 18 million SNPs). We thus do not recommend running PolyFun-pred with a reduced SNP set. We have updated the *Simulations with in-sample LD* subsection of the Results section (p.6, citing Supplementary Table 1) and the Discussion section (p.15) to clarify this point.

We note that we applied P+T using 18 million MAF>0.1% SNPs, similar to PolyPred. However, using P+T with a restricted SNP set is very unlikely to improve its performance, because it selects a single SNP from each LD block, and is thus very unlikely to be adversely affected by a denser SNP set.

(c) What is the impact of summary LD data on PRS accuracy?

Summary LD can impact PRS accuracy via (i) SNP density, (ii) sample size of the LD reference panel, and (iii) similarity between the ancestry of the LD reference panel and the ancestry of the training samples from which summary statistics are analyzed to compute PRS. We discuss each of these factors in turn.

(i) SNP density.

As noted in (b) above, SBayesR-2.8M was less accurate than SBayesR (significantly so for Africans), and there is little reason to believe that further expanding the SNP set used to compute summary LD would improve the performance of SBayesR. On the other hand, PolyFun-pred relies on fine-mapping, which can be severely compromised by using a reduced SNP set. To demonstrate this, we performed simulations in which we trained PolyFun-pred using only 1.2 million HapMap3 SNPs (instead of 18 million MAF>0.1% SNPs) and obtained substantially and significantly reduced accuracy (-11% relative-R^2 in non-British Europeans vs. using all 18 millions SNPs). We thus do not recommend running PolyFun-pred with a reduced SNP set. We have updated the *Simulations with in-sample LD* subsection of the Results section (p.6) and the Discussion section (p.15) to clarify this point.

(ii) sample size of the LD reference panel.

Larger LD reference panels improve PRS accuracy because they enable more accurate LD estimates. We performed three new experiments to verify this:

First, we ran PolyFun-pred using summary LD from UK10K (N=3,567) in real trait analyses, or various subsets of UK10K (N=489-3,567) in simulations, instead of summary LD from UK Biobank (337K LD reference samples) (also see (iii) below for analyses using summary LD from 1000 Genomes Europeans). In real trait analyses, PolyFun-pred using the UK10K LD reference panel suffered a substantial and significant loss of accuracy compared to using in-sample LD from UK Biobank British individuals (N=337K) (−86% relative-R^2). Results are reported in Supplementary Table 4, cited in the *Analysis of 4 UK Biobank populations using UK Biobank British training data* subsection of the Results section (p. 8). In simulations, the accuracy of PolyFun-pred decreased as we subsampled smaller and smaller subsets of UK10K, with a 92% smaller relative-R^2 when using only N=489 UK10K individuals (the same sample size as the 1000 Genomes project Europeans). Results are reported in Supplementary Table 1, cited in the *Simulations with reference LD* subsection of the Supplementary Note (p.6), which is cited in the *Simulations with in-sample LD* subsection of the Results section (p.5).

Second, we ran SBayesR using summary LD from various subsets of UK10K (N=489-3,567) in real trait analyses, or from UK10K (N=3,567) in simulations (also see (iii) below for analyses using summary LD from 1000 Genomes Europeans). In real trait analyses, SBayesR accuracy using the full UK10K LD reference panel was very similar to and statistically indistinguishable from the accuracy obtained using the LD reference panels provided by the authors of SBayesR (N=50K UK Biobank British individuals). However, the relative-R^2 was 24% smaller when using only N=489 UK10K individuals. Results are reported in Supplementary Table 4, cited in the *Analysis of 4 UK Biobank populations using UK Biobank British training data* subsection of the Results section (p.8). In simulations, we observed a smaller but statistically significant loss of accuracy when using a UK10K LD reference panel. Results are reported in Supplementary Table 1, cited in the *Simulations with reference LD* subsection of the Supplementary Note (p.6), which is cited in the *Simulations with in-sample LD* subsection of the Results section (p.5).

Third, we ran PRS-CS using summary LD from the 1000 Genomes project Europeans (N=489), in both simulations and real trait analyses (we could not use summary LD from UK10K because PRS-CS does not allow computing LD matrices). In real trait analyses, the accuracy of PRS-CS was very similar to the accuracy obtained using in-sample LD from UK Biobank British individuals (N=337K), with no significant differences. Results are reported in Supplementary Table 4, cited in the *Analysis of 4 UK Biobank populations using UK Biobank British training data* subsection of the Results section (p.8). Interestingly, in simulations, the relative-R^2 of PRS-CS using summary LD from the 1000 Genomes project Europeans was up to 42% smaller than when using in-sample LD from UK Biobank British, suggesting that the effect of LD reference panel sample size on PRS-CS accuracy is sensitive to the underlying genetic architecture. Results are reported in Supplementary Table 1, cited in the *Simulations with reference LD* subsection of the Supplementary Note (p.6), which is cited in the *Simulations with in-sample LD* subsection of the Results section (p.5).

(iii) similarity between the ancestry of the LD reference panel and the ancestry of the training samples from which summary statistics are analyzed to compute PRS.

The LD reference panel should ideally have the same ancestry as the training samples from which summary statistics are analyzed to compute PRS. In our study, the training samples consist of up to 337K British Europeans from UK Biobank, and the LD reference panel in our primary analyses consists of the same set of 337K British Europeans.

It is challenging to precisely quantify the impact of ancestry mismatch, as we do not currently have access to a large (e.g. *N*>3,000) sequenced non-British European sample with which to explore the impact of ancestry mismatch; analyses of smaller LD reference samples would likely be dominated by the impact of sample size of the LD reference panel (see (ii) above).

Nevertheless, we attempted to evaluate the effect of LD mismatch by running PolyFun-pred, SBayesR, and PRS-CS using an LD reference panel consisting of *N*=489 European individuals from the 1000 Genomes project, in both simulations and real trait analyses. We obtained reduced accuracy in simulations for all methods, and in real trait analyses for PolyFun-pred and SBayesR; the difference was not statistically significant for PRS-CS in real trait analyses. In the case of SBayesR, we determined in real trait analyses that this loss of accuracy primarily stems from LD mismatch (rather than reduced sample size) by repeating the analysis with *N*=489 individuals from UK10K, for which we obtained significantly improved accuracy. On the other hand, results for PolyFun-pred were significantly less accurate even when using UK10K individuals in real trait analyses, suggesting that the less accurate PolyFun-pred results are primarily driven by reduced sample size. Simulation results are reported in Supplementary Table 1, cited in the *Simulations with reference LD* subsection of the Supplementary Note (p.6), which is cited in the *Simulations with in-sample LD* subsection of the Results section (p.5). Real trait analysis results are reported in Supplementary Table 4, cited in the *Analysis of 4 UK Biobank populations using UK Biobank British training data* subsection of the Results section (p.8).

5. Also in the simulation the authors describe “Using UK10K was almost as accurate as using in-sample LD, but using 1000 Genomes Europeans was hugely inaccurate, leading to prediction

accuracy even lower than that of P+T (Supplementary Table 1), confirming the importance of using a large (population-matched) LD reference panel to compute PRS41 (on the other hand, using UK10K LD is not recommended when using PolyFun for fine-mapping29 due to concerns about false-positive fine-mapped SNPs, which are not a primary concern when computing PRS).” Did you also try a downsampled set of UK10K of the same size as the 1000 Genomes project Europeans (N=-489) to confirm whether it is primarily sample size or the underlying LD patterns that are driving the results? I expect the 1000 Genomes Europeans are of more diverse ancestry that the 3500 UK10K. So, is the sample size issue creating more smaller LD blocks than the UK10K? Would it look more similar if both are less than 500 people to estimate the LD?

(also see part (c).(ii) of response to Reviewer #1 Comment 4)

We agree that this is a valuable experiment. We have performed new simulations and real data analysis to investigate the effect of using LD summary from either the full or a reduced subset of UK10K on the accuracy of PolyFun-pred and SBayesR (we did not perform this experiment for PRS-CS because the PRS-CS software does not allow computing custom-tailored LD matrices).

First, we ran PolyFun-pred using summary LD from UK10K (N=3,567) in real trait analyses, or various subsets of UK10K (N=489-3,567) in simulations, instead of summary LD from UK Biobank (337K LD reference samples). In real trait analyses, PolyFun-pred using the UK10K LD reference panel suffered a substantial and significant loss of accuracy compared to using in-sample LD from UK Biobank British individuals (N=337K) (−86% relative-R^2). Results are reported in Supplementary Table 4, cited in the *Analysis of 4 UK Biobank populations using UK Biobank British training data* subsection of the Results section (p.8). In simulations, the accuracy of PolyFun-pred decreased as we subsampled smaller and smaller subsets of UK10K, with a 92% smaller relative-R^2 when using only N=489 UK10K individuals (the same sample size as the 1000 Genomes project Europeans). Results are reported in Supplementary Table 1, cited in the *Simulations with reference LD* subsection of the Supplementary Note (p.6), which is cited in the *Simulations with in-sample LD* subsection of the Results section (p.5).

Second, we ran SBayesR using summary LD from various subsets of UK10K (N=489-3,567) in real trait analyses, or from UK10K (N=3,567) in simulations. In real trait analyses, SBayesR accuracy using the full UK10K LD reference panel was very similar to and statistically indistinguishable from the accuracy obtained using the LD reference panels provided by the authors of SBayesR (N=50K UK Biobank British individuals). However, the relative-R^2 was 24% smaller when using only N=489 UK10K individuals. Results are reported in Supplementary Table 4, cited in the *Analysis of 4 UK Biobank populations using UK Biobank British training data* subsection of the Results section (p.8). In simulations, we observed a smaller but statistically significant loss of accuracy when using a UK10K LD reference panel. Results are reported in

Supplementary Table 1, cited in the *Simulations with reference LD* subsection of the Supplementary Note (p.6), which is cited in the *Simulations with in-sample LD* subsection of the Results section (p.5).

6. The compute time differences are quite large (2.8 minutes versus 668 minutes). What is causing the 300-times longer compute? How much of an impact will this have on the ability to use PolyPred in real-world scenarios?

The reviewer has raised 2 related questions: (a) what is causing the 300-times larger compute time for PolyPred vs. SBayesR?; and (b) what is the impact of the larger compute time of PolyPred in real-world settings? We address each question in turn.

(a) What is causing the 300-times larger compute time for PolyPred vs. SBayesR?

The larger compute time of PolyPred (and its summary statistic-based analogues) is dominated by the PolyFun-pred component, which is computationally intensive because (i) PolyFun-pred performs fine-mapping, which is a more computationally intensive task than other approaches to computing PRS coefficients (e.g. computing posterior mean tagging effect sizes under an assumed prior, as in SBayesR); and (ii) PolyFun-pred analyzes a large number of SNPs, e.g. 18 million SNPs in UK Biobank training data and 8.1 million SNPs in ENGAGE training data (vs. 1.2 million SNPs for SBayesR). We note that running SBayesR using 18 million SNPs is not computationally feasible, but we determined that running SBayesR using 2.8 million SNPs (SBayesR-2.8M) is less accurate than SBayesR (significantly so for Africans) (see part (b) of response to Reviewer #1 Comment 4). We further note that all of the above points pertain to the time required to train the PRS model, and not the time required to apply the PRS model to compute predictions in target samples (which is extremely small). We have updated the *Simulations with in-sample LD* subsection of the Results section (p.6) and the Discussion section (p.15) to clarify these points.

(b) What is the impact of the larger compute time of PolyPred in real-world settings?

We anticipate that the larger compute time of PolyPred will have little impact in real-world settings. The relatively computationally expensive step of training the PRS model is performed only once (and can easily be parallelized across loci), whereas the frequently repeated set of applying the PRS model to compute predictions in test samples is computationally fast (with the same computational cost for PolyPred vs. other methods). We have updated the Discussion section (p.15) to clarify this point.

7. At the end of the results section, the authors say “We emphasize that efforts to assess the benefit of incorporating non-European training data should analyze non-European training data

from a cohort that is distinct from the target cohort, otherwise results may be inflated due to cohort effects." This is perhaps one of the most critically important points of the paper. Does this mean that the use of very large, trans-ancestry meta-analyses to estimate the weights in the training data would be optimal as long as the target cohort is not included? Did you evaluate any of the traits using this approach where a meta-analysis of many studies and/or ancestry groups is the basis of the training data?

The reviewer has raised 3 related questions: (a) is there an "optimal" way to choose a training cohort that is distinct from the target cohort?; (b) can PolyPred be applied to training data consisting of a meta-analysis of many studies?; and (c) can PolyPred be applied to training data consisting of a meta-analysis of different ancestry groups? We address each question in turn.

(a) Is there an "optimal" way to choose a training cohort that is distinct from the target cohort?

We believe there is no single "optimal" way to choose a training cohort that is distinct from the target cohort. In particular, training sample size is a critical factor impacting PRS accuracy, and choices that maximize training sample size will depend on the particular disease/trait analyzed. We have updated the Discussion section (p.14) to clarify this point. In (b) and (c) below, we discuss two specific strategies mentioned by the reviewer.

(b) Can PolyPred be applied to training data consisting of a meta-analysis of many studies?
In our previously submitted manuscript, we stated that PolyPred-S (which linearly combines PolyFun-pred and SBayesR) can be applied instead of PolyPred (which linearly combines PolyFun-pred and BOLT-LMM) when only summary statistics are available; we further stated that a large (N>3K) LD reference panel should be used. Training data consisting of a meta-analysis of many studies is an example of this scenario. We agree with the reviewer's feedback that this is a very important scenario to consider, and have thus performed a new analysis of real traits pertaining to this scenario (also see response to part (b) of Reviewer #2 Comment 2). Our results are summarized in Figure 5 (and the new Table 2) and are detailed below.

We used summary statistics from the European Network for Genetic and Genomic Epidemiology (ENGAGE) consortium (Budin-Ljøsne *et al*. 2014 Eur J Hum Genet, ref. 68) to train several methods on four traits (BMI, waist-hip-ratio (adjusted for BMI), total cholesterol, and triglycerides), and evaluated the prediction accuracy using the same four UK Biobank populations analyzed throughout our manuscript (Non-British Europeans, South-Asians, East-Asians, and Africans). We selected this particular meta-analysis because it includes a dense set of 8.1 million imputed SNPs, which enables fine-mapping. For each method, we used the same LD reference panel used in the other primary analyses, based on UK Biobank British individuals; we emphasize that unlike the other primary analyses in this manuscript, the LD reference panel was misspecified, because it was not based on in-sample LD. We excluded BOLT-LMM from this analysis (because it cannot use summary statistics) and included PRS-CS, following our decision to include PRS-CS as a main method in the manuscript. In the following, we denote PolyPred-P as the linear combination of PolyFun-pred and PRS-CS, and PolyPred-S as the linear combination of PolyFun-pred and SBayesR.

The results are summarized in our new Figure 5. Briefly, PolyPred-P was generally the most accurate method, and PRS-CS outperformed SBayesR, with a significant improvement for non-British Europeans and Africans (unlike when using UK Biobank training data, where SBayesR outperformed PRS-CS). However, the differences between the methods were mostly not statistically significant due to large standard errors.

In detail, the average relative-R^2 in Non-British Europeans was 0.045 for PolyPred-P, 0.044 for PolyPred-S, 0.039 for PRS-CS, 0.033 for SBayesR, and 0.022 for P+T. For African-ancestry individuals (obtaining the lowest prediction accuracy under all methods), the average relative-R^2 was 0.015 for PolyPred-P, 0.008 for PolyPred-S, 0.013 for PRS-CS, 0.004 for SBayesR, and 0.010 for P+T. We reiterate that the differences between the methods were mostly not statistically significant due to moderately large standard errors (Figure 5), and thus caution should be exercised in their interpretation. Nevertheless, these results demonstrate that combining PolyFun-pred with another method can be beneficial when analyzing summary statistics from a meta-analysis of many studies. We further note that PRS-CS outperforms SBayesR in the presence of a misspecified LD reference panel, consistent with a previous report (Pain *et al.* 2021 PloS Genet, ref. 71). We have added a new *Analysis of 4 UK Biobank populations using ENGAGE meta-analysis training data* subsection of the Results section (p.10, citing Figure 5) to include these results.

We have also performed new simulations and real trait analyses of UK Biobank training data using reference LD; see part (c) (ii) of response to Reviewer #1 Comment 4 and response to Reviewer #1 Comment 5.

(c) Can PolyPred be applied to training data consisting of a meta-analysis of different ancestry groups?

One of the main conclusions of our work is that leveraging training data from different ancestry groups (e.g. different continental ancestries) improves PRS in diverse populations. However, we recommend against using training data consisting of a traditional fixed-effect meta-analysis of GWAS data from different ancestry groups, for two reasons: (i) fixed-effect meta-analysis implies that European training samples and training samples from the non-European target population would receive equal weight, whereas our work shows that the latter should receive higher weight in order to maximize PRS accuracy; and (ii) it may be challenging to construct an LD reference panel whose ancestry matches the ancestry of the meta-analysis of different ancestry groups. When possible, it would be preferable to separately incorporate European training data and training data from the non-European target population, with appropriate LD reference panels. We have updated the Discussion section (p.15) to clarify this point.

8. In the discussion, the authors state “Finally, PRS may implicitly capture GxE interactions, which may not be transferable across cohorts or ancestries27,75.” This is an interesting point. But why this would be the case is not clear based on any of the analyses provided here. Can this be explained further? How are the GxE effects being captured by PRS and more importantly, how can we determine whether they are transferable across ancestry?

We believe that the most interesting case is that of a GxE effect in which the GxE effect is shared across ancestries but the (average) value of E differs across ancestries. If E (and GxE) is unmodeled, G effects will (appear to) differ across ancestries; this is a possible explanation for cross-population genetic correlations significantly less than 1 for some diseases/traits, which are well-documented (Martin et al. 2019 Nat Genet (ref. 13), Shi et al. 2021 Nat Commun (ref. 30)). This is one of the motivations for leveraging training data from non-European target populations. However (if E is unmodeled), it is difficult to distinguish this scenario from the scenario of different G effects for other reasons. We have updated the Discussion section (p.14) to clarify these points.

9. In the methods, where do you get the 187 functional annotations for Polyfun-pred?

The 187 functional annotations for PolyFun-pred were previously described in Weissbrod et al. 2020 Nat Genet (ref. 35), which introduced the PolyFun method. We have updated the *PolyPred and its summary statistic-based analogues* subsection of the Methods section (p.18) to clarify that the 187 functional annotations were previously described in Weissbrod et al. 2020, to add text summarizing the content of the 187 functional annotations (10 common MAF bins (MAF≥0.05); 10 low-frequency MAF bins (0.05>MAF≥0.001); 6 LD-related annotations for common SNPs (levels of LD, predicted allele age, recombination rate, nucleotide diversity, background selection statistic, CpG content); 5 LD-related annotations for low-frequency SNPs; 40 binary functional annotations for common SNPs; 7 continuous functional annotations for common SNPs; 40 binary functional annotations for low-frequency SNPs; 3 continuous functional annotations for low-frequency SNPs; and 66 annotations constructed via windows around other annotations), and to cite a new Supplementary Table 11 listing the 187 functional annotations.

10. Throughout the paper, the authors use the term “ethnicity”, including in the figures. Is “ethnicity” the correct word? It seems as though the focus is really on genetic ancestry, which is different from ethnicity. As we try to move as a field toward removing racism and bias, it seems like being more precise with our words, such as the use of the word “ancestry” when we are truly using genetics to define groups, seems more appropriate. If I am misunderstanding the way that groups are defined herein, and ethnicity is more accurate, please keep ethnicity but consider adding an explanation in the paper as to why that word is being used.

We agree with the reviewer that our analyses pertain to ancestry and not to ethnicity. We have changed all instances of “ethnicity” (resp. “trans-ethnic”) to “ancestry” (resp. “cross-population”).

Reviewer #2: (numbers added to reviewer comments)

1. First of all, this manuscript is quite technical and it took me some time to understand its flow. An additional complication is that not all methods are usable without individual level data, which further complicates the choice of methodology. These points are discussed in various sections of the manuscript, but a key comment would be whether the authors would consider a flowchart of some sort that summarises how the various methods fit together, with or without the availability of individual level data. I am aware that saying that a manuscript is hard to follow is a typically unhelpful reviewer's comment, but I think it's fair to say that the ratio text:display items is high, and additional help to understand how it all fits would help. More sub-headers would perhaps also be helpful, as some sections are very lengthy and cover a lot of ground.

We agree that the manuscript is quite technical and thus difficult to follow. The reviewer has suggested that we (a) add a flowchart summarizing the various methods, with or without the availability of individual-level data, (b) consider adding additional display items, and (c) add more sub-headings. We thank the reviewer for the suggestions, and address each suggestion in turn.

(a) add a flowchart summarizing the various methods, with or without the availability of individual-level data.

We agree and have added this flowchart: Figure 2, cited in the *Overview of methods* subsection of the Results section (p.4) and the Discussion section (p.13).

(b) consider adding additional display items.

We agree and have added 4 new display items:

A new Figure 1, cited in the *Overview of methods* subsection of the Results section (p.4), providing a schematic overview of PolyPred (and PolyPred-S and PolyPred-P).

A new Figure 2, cited in the *Overview of methods* subsection of the Results section (p.4) and the Discussion section (p.13), providing a flowchart summarizing the various methods (see (a) above).

A new Table 2, cited in the *Overview of methods* subsection of the Results section (p.4) and the Discussion section (p.13), providing a summary of the relative performance of constituent PRS methods under various settings, and links to the corresponding Figures/Tables.

A new Figure 5, cited in the new *Analysis of 4 UK Biobank populations using ENGAGE meta-analysis training data* subsection of the Results section (p.10), providing the results of the new ENGAGE analysis (see part (b) of response to Reviewer #1 Comment 7 and part (b) of response to Reviewer #2 Comment 2).

(c) add more sub-headings.

We have split the *Simulations* subsection of the Results section into two subsections: the *Simulations with in-sample LD* subsection of the Results section, and a new Supplementary Note subsection titled *Simulations with reference LD* (see part (c) (ii) of response to Reviewer #1 Comment 4 and response to Reviewer #1 Comment 5). We note that the *Simulations* subsection of the Results section was the longest subsection (and more simulations have been added). We are open to moving the *Simulations with reference LD* subsection of the Supplementary Note to the Results section if editors and/or reviewers express a strong preference, but we believe this would likely be incompatible with the Nature Genetics word count limit.

We have also added a new Results subsection titled *Analysis of 4 UK Biobank populations using ENGAGE meta-analysis training data* (p.10), citing a new Figure 5 (see part (b) of response to Reviewer #1 Comment 7 and part (b) of response to Reviewer #2 Comment 2).

We have elected not to add sub-subsection headings, but we are open to adding these if the editors and/or reviewers express a strong preference.

2. Among these complications, the requirement for individual level data is in fact quite critical, as one may argue that the comparisons are somewhat academic given that the use of summary statistics typically yields much larger training sets. It does not mean that the insights from this paper aren't important, but I would like this individual level data requirement to be made clearer (for example in the introduction).

The reviewer has raised 2 related concerns: (a) the dependence of the PolyPred method on individual-level data should be more strongly emphasized; and (b) the comparisons are somewhat academic given the importance of summary statistic data sets. We address each concern in turn.

(a) The dependence of the PolyPred method on individual-level data should be more strongly emphasized.

We have modified the Abstract (p.1), Introduction section (p.2) and Discussion section (p.14) to more strongly emphasize the dependence of PolyPred on individual-level data, while also noting that the same framework can be applied to summary statistic data sets (PolyPred-S and PolyPred-P; see (b)).

(b) The comparisons are somewhat academic given the importance of summary statistic data sets.

We agree that the analysis of summary statistic data sets is important (also see part (b) of response to Reviewer #1 Comment 7).

In our previously submitted manuscript, we stated that PolyPred-S (which linearly combines PolyFun-pred and SBayesR) can be applied instead of PolyPred (which linearly combines PolyFun-pred and BOLT-LMM) when only summary statistics are available; we further stated that a large (*N*>3K) LD reference panel should be used. Training data consisting of a meta-analysis of many studies is an example of this scenario. We agree with the reviewer's feedback that this is a very important scenario to consider and have thus performed a new analysis of real traits pertaining to this scenario (also see response to part (b) of Reviewer #1 Comment 7). Our results are summarized in Figure 5 (and the new Table 2) and are detailed below.

We used summary statistics from the European Network for Genetic and Genomic Epidemiology (ENGAGE) consortium (Budin-Ljøsne *et al*. 2014 Eur J Hum Genet, ref. 68) to train several methods on four traits (BMI, waist-hip-ratio (adjusted for BMI), total cholesterol, and triglycerides), and evaluated the prediction accuracy using the same four UK Biobank populations analyzed throughout our manuscript (Non-British Europeans, South-Asians, East-Asians, and Africans). We selected this particular meta-analysis because it includes a dense set of 8.1 million imputed SNPs, which enables fine-mapping. For each method, we used the same LD reference panel used in the other primary analyses, based on UK Biobank British individuals; we emphasize that unlike the other primary analyses in this manuscript, the LD reference panel was misspecified, because it was not based on in-sample LD. We excluded BOLT-LMM from this analysis (because it cannot use summary statistics) and included PRS-CS, following our decision to include PRS-CS as a main method in the manuscript. In the following, we denote PolyPred-P as the linear combination of

PolyFun-pred and PRS-CS, and PolyPred-S as the linear combination of PolyFun-pred and SBayesR.

The results are summarized in our new Figure 5. Briefly, PolyPred-P was generally the most accurate method, and PRS-CS outperformed SBayesR, with a significant improvement for non-British Europeans and Africans (unlike when using UK Biobank training data, where SBayesR outperformed PRS-CS). However, the differences between the methods were mostly not statistically significant due to large standard errors.

In detail, the average relative-R^2 in Non-British Europeans was 0.045 for PolyPred-P, 0.044 for PolyPred-S, 0.039 for PRS-CS, 0.033 for SBayesR, and 0.022 for P+T. For African-ancestry individuals (obtaining the lowest prediction accuracy under all methods), the average relative-R^2 was 0.015 for PolyPred-P, 0.008 for PolyPred-S, 0.013 for PRS-CS, 0.004 for SBayesR, and 0.010 for P+T. We reiterate that the differences between the methods were mostly not statistically significant due to moderately large standard errors (Figure 5), and thus caution should be exercised in their interpretation. Nevertheless, these results demonstrate that combining PolyFun-pred with another method can be beneficial when analyzing summary statistics from a meta-analysis of many studies. We further note that PRS-CS outperforms SBayesR in the presence of a misspecified LD reference panel, consistent with a previous report (Pain *et al.* 2021 PloS Genet, ref. 71). We have added a new *Analysis of 4 UK Biobank populations using ENGAGE meta-analysis training data* subsection of the Results section (p.10, citing Figure 5) to include these results.

We have also performed new simulations and real trait analyses of UK Biobank training data using reference LD; see part (c) (ii) of response to Reviewer #1 Comment 4 and response to Reviewer #1 Comment 5.

3. It is obviously another layer of complication but I would like to see what Polypred-S is doing in the main display items (figures 2-4).

We agree, and have updated Table 1, Figure 2 (new), Figure 3 (formerly Figure 1), Figure 4 (formerly Figure 2), Figure 5 (new), Figure 6 (formerly Figure 3) and Figure 7 (formerly Figure 4) to include PolyPred-S (which uses SBayesR) and PolyPred-P (which uses PRS-CS) (we have added PRS-CS as a main method throughout our revised manuscript; see response to Reviewer #2 Comment 5). In detail:

Figure 3 (formerly Figure 1) now includes PolyPred-S and PolyPred-P, and we have updated the *Simulations with in-sample LD* subsection of the Results section (p. 4-5) accordingly.

Figure 4 (formerly Figure 2) now includes PolyPred-S and PolyPred-P, and we have updated the *Analysis of 4 UK Biobank populations using UK Biobank British training data* subsection of the Results section (p. 6-9) accordingly.

Figure 6 (formerly Figure 3) now includes PolyPred-S and PolyPred-P, and we have updated the *Analysis of Biobank Japan and Uganda-APCDR cohorts* subsection of the Results section (p.10-12) accordingly.

Figure 7 (formerly Figure 4) now includes PolyPred-S (and PolyPred-S+) and PolyPred-P (and PolyPred-P+), and we have updated the *Analysis of UK Biobank East Asians using UK Biobank British and Biobank Japan training data* subsection of the Results section (p.12-13) accordingly.

We have also included PolyPred-S and PolyPred-P in Table 1 and in the new Figure 2, cited in the *Overview of methods* subsection of the Results section (p.4) (see part (a) of response to Reviewer #2 Comment 1), and in the new Figure 5, cited in the new *Analysis of 4 UK Biobank populations using ENGAGE meta-analysis training data* subsection of the Results section (p.10) (see part (b) of response to Reviewer #1 Comment 7 and part (b) of response to Reviewer #2 Comment 2).

4. last major point relates to the general methodology of linearly combining different PRS. Fundamentally, BOLT-LMM and Polypred-fun are capturing much of the same signal, so there is some double counting and some single counting of signals, which has to be sub-optimum. I would argue that some form of signal subtraction would make more sense. I wonder if the authors could comment on that point. Dealing with this issue properly is probably beyond the scope of this manuscript, but I think it would be helpful to consider what could be done.

We agree that prediction accuracy could in principle be improved if it were possible to decompose the PolyFun-pred and BOLT-LMM predictors into shared and non-shared components, to improve upon double counting of shared components vs. single counting of non-shared components.

We have updated the Discussion section (p.15) and added a subsection called *Decomposing the PolyFun-pred and BOLT-LMM predictors into shared and non-shared components* to the Supplementary Note (p. 8) (cited in the Discussion section, p.15) to note this potential for improvement. We are not currently aware of any way to decompose the BOLT-LMM and PolyFun-pred predictors into shared and non-shared components. (We did explore the use of different mixing weights in different segments of the genome, but did not obtain promising results. Given the large number of secondary analyses, we prefer not to include these findings in the manuscript, unless the editors and/or reviewers express a strong preference).

5. Lastly, on a more minor point, the authors speak well of the SBayesR results outperforming PRS-CS and LDPred. This puzzles me, as I must say I never managed to get SBayesR to run properly. My struggles are echoed by the recent bioRxiv manuscript from Pain, Lewis and colleagues which I am sure the authors are aware of (https://doi.org/10.1101/2020.07.28.224782). I'd be keen to understand why such stark differences are being observed.

The reviewer is correct that Pain et al. demonstrated that PRS-CS outperforms SBayesR in the presence of mismatch between the GWAS sample and the LD reference panel (Pain *et al*. 2021 PLoS Genet, ref. 71). We performed a new analysis of real traits to investigate this (also see part (b) of response to Reviewer #1 Comment 7, part (b) of response to Reviewer #1 Comment 2, and part (b) of response to Reviewer #2 Comment 2). Our results are summarized in Figure 5 (and the new Table 2) and are detailed below.

We used summary statistics from the European Network for Genetic and Genomic Epidemiology (ENGAGE) consortium (Budin-Ljøsne *et al*. 2014 Eur J Hum Genet, ref. 68) to train several methods on four traits (BMI, waist-hip-ratio (adjusted for BMI), total cholesterol, and triglycerides), and evaluated the prediction accuracy using the same four UK Biobank populations analyzed throughout our manuscript (Non-British Europeans, South-Asians, East-Asians, and Africans). We selected this particular meta-analysis because it includes a dense set of 8.1 million imputed SNPs, which enables fine-mapping. For each method, we used the same LD reference panel used in the other primary analyses, based on UK Biobank British individuals; we emphasize that unlike the other primary analyses in this manuscript, the LD reference panel was misspecified, because it was not based on in-sample LD. We excluded BOLT-LMM from this analysis (because it cannot use summary statistics) and included PRS-CS, following our decision to include PRS-CS as a main method in the manuscript. In the following, we denote PolyPred-P as the linear combination of PolyFun-pred and PRS-CS, and PolyPred-S as the linear combination of PolyFun-pred and SBayesR.

The results are summarized in our new Figure 5 (and the new Table 2). Briefly, PolyPred-P was generally the most accurate method, and PRS-CS outperformed SBayesR, with a significant improvement for non-British Europeans and Africans (unlike when using UK Biobank training data, where SBayesR outperformed PRS-CS). However, the differences between the methods were mostly not statistically significant due to large standard errors.

In detail, the average relative-R^2 in Non-British Europeans was 0.045 for PolyPred-P, 0.044 for PolyPred-S, 0.039 for PRS-CS, 0.033 for SBayesR, and 0.022 for P+T. For African-ancestry individuals (obtaining the lowest prediction accuracy under all methods), the average relative-R^2 was 0.015 for PolyPred-P, 0.008 for PolyPred-S, 0.013 for PRS-CS, 0.004 for SBayesR, and 0.010 for P+T. We reiterate that the differences between the methods were mostly not statistically significant due to moderately large standard errors (Figure 5), and thus caution should be exercised in their interpretation. Nevertheless, these results demonstrate that combining PolyFun-pred with another method can be beneficial when analyzing summary statistics from a meta-analysis of many studies. We further note that PRS-CS outperforms SBayesR in the presence of a misspecified LD reference panel, consistent with Pain *et al.* 2021 PloS Genet (ref. 71). We have added a new *Analysis of 4 UK Biobank populations using ENGAGE meta-analysis training data* subsection of the Results section (p.10, citing Figure 5) to include these results.

We have also performed new simulations and real trait analyses of UK Biobank training data using reference LD; see part (c) (ii) of response to Reviewer #1 Comment 4 and response to Reviewer #1 Comment 5.

In light of the fact that PRS-CS outperforms SBayesR in some settings, we have added PRS-CS and PolyPred-P as main methods throughout our revised manuscript, including Table 1, Table 2 (new), Figure 2 (new), Figure 3 (formerly Figure 1), Figure 4 (formerly Figure 2), Figure 5 (new), Figure 6 (formerly Figure 3) and Figure 7 (formerly Figure 4).

Decision Letter, first revision:

9th Nov 2021

Dear Dr. Weissbrod,

Thank you for submitting your revised manuscript "Leveraging fine-mapping and non-European training data to improve cross-population polygenic risk scores" (NG-A56478R1). It has now been seen by the original referees and their comments are below. The reviewers find that the paper has improved in revision, and therefore we'll be happy in principle to publish it in Nature Genetics, pending minor revisions to satisfy the referees' final requests and to comply with our editorial and formatting guidelines.

If the current version of your manuscript is in a PDF format, please email us a copy of the file in an editable format (Microsoft Word or LaTex)-- we can not proceed with PDFs at this stage.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements soon. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Thank you again for your interest in Nature Genetics Please do not hesitate to contact me if you have any questions.

Sincerely,

Wei

Wei Li, PhD
Senior Editor
Nature Genetics
New York, NY 10004, USA
www.nature.com/ng

Reviewer #1 (Remarks to the Author):

The authors have done a tremendous job responding to the previous reviews. I think the work is thorough, clear, and very well done. I have no further revision suggestions.

Reviewer #2 (Remarks to the Author):

I would like to thank the authors for what is clearly a detailed reply and an improved manuscript. The added figures and text help clarify the work, and all comments have been taken very seriously.

One minor remaining comment: I was going through the supplemental tables trying to interpret the performance of the method with my current benchmarks in mind, and all my personal references for binary traits (CVD, T2D...) are based on AUC or OR per SD. I would find it helpful to convert, for binary traits, the r2 statements made in these tables into the more broadly used AUC statistic. It is not essential, but it would help I think.

To conclude, and as mentioned in my initial review, I do find the overall body of work impressive, and I definitely think it should be published. It does remain very technical, and perhaps longer than most NatGen papers, but I do not think this is a major issue. In any case, this is more an editorial than a review issue. From a technical perspective, the paper is sound, detailed and valuable to the field.

Final Decision Letter:

25th Feb 2022

Dear Dr. Weissbrod,

I am delighted to say that your manuscript "Leveraging fine-mapping and multi-population training data to improve cross-population polygenic risk scores" has been accepted for publication in an upcoming issue of Nature Genetics.

Over the next few weeks, your paper will be copyedited to ensure that it conforms to Nature Genetics style. Once your paper is typeset, you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required.

After the grant of rights is completed, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately.

You will not receive your proofs until the publishing agreement has been received through our system.

Due to the importance of these deadlines, we ask that you please let us know now whether you will be difficult to contact over the next month. If this is the case, we ask you provide us with the contact information (email, phone and fax) of someone who will be able to check the proofs on your behalf, and who will be available to address any last-minute problems.

Your paper will be published online after we receive your corrections and will appear in print in the next available issue. You can find out your date of online publication by contacting the Nature Press Office (press@nature.com) after sending your e-proof corrections. Now is the time to inform your Public Relations or Press Office about your paper, as they might be interested in promoting its publication. This will allow them time to prepare an accurate and satisfactory press release. Include your manuscript tracking number (NG-A56478R2) and the name of the journal, which they will need when they contact our Press Office.

Before your paper is published online, we shall be distributing a press release to news organizations worldwide, which may very well include details of your work. We are happy for your institution or funding agency to prepare its own press release, but it must mention the embargo date and Nature Genetics. Our Press Office may contact you closer to the time of publication, but if you or your Press Office have any enquiries in the meantime, please contact press@nature.com.

Acceptance is conditional on the data in the manuscript not being published elsewhere, or announced in the print or electronic media, until the embargo/publication date. These restrictions are not intended to deter you from presenting your data at academic meetings and conferences, but any enquiries from the media about papers not yet scheduled for publication should be referred to us.

Please note that <i>Nature Genetics</i> is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. <a href="https://www.springernature.com/gp/open-research/transformative-journals"> Find out more about Transformative Journals</a>

<B>Authors may need to take specific actions to achieve <a href="https://www.springernature.com/gp/open-research/funding/policy-compliance-faqs"> compliance</a> with funder and institutional open access mandates. For submissions from January 2021, if your research is supported by a funder that requires immediate open access (e.g. according to <a href="https://www.springernature.com/gp/open-research/plan-s-compliance">Plan S principles</a>) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route our standard licensing terms will need to be accepted, including our <a href="https://www.springernature.com/gp/open-research/policies/journal-policies">self-archiving policies</a>. Those standard licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

Please note that Nature Research offers an immediate open access option only for papers that were first submitted after 1 January, 2021.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

If you have posted a preprint on any preprint server, please ensure that the preprint details are updated with a publication reference, including the DOI and a URL to the published version of the article on the journal website.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

As soon as your article is published, you will receive an automated email with your shareable link.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

An online order form for reprints of your paper is available at <a href="https://www.nature.com/reprints/author-reprints.html">https://www.nature.com/reprints/author-reprints.html</a>. Please let your coauthors and your institutions' public affairs office know that they are also welcome to order reprints by this method.

If you have not already done so, we invite you to upload the step-by-step protocols used in this manuscript to the Protocols Exchange, part of our on-line web resource, natureprotocols.com. If you complete the upload by the time you receive your manuscript proofs, we can insert links in your article that lead directly to the protocol details. Your protocol will be made freely available upon publication of your paper. By participating in natureprotocols.com, you are enabling researchers to more readily reproduce or adapt the methodology you use. Natureprotocols.com is fully searchable, providing your protocols and paper with increased utility and visibility. Please submit your protocol to https://protocolexchange.researchsquare.com/. After entering your nature.com username and password you will need to enter your manuscript number (NG-A56478R2). Further information can be found at https://www.nature.com/nprot/.

Sincerely,

Wei Li, PhD
Senior Editor
Nature Genetics
New York, NY 10004, USA
www.nature.com/ng