nature portfolio

Peer Review Information

Journal: Nature Genetics
Manuscript Title: Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity
Corresponding author name(s): Dr Steven Gazal

Reviewer Comments & Decisions:

Decision Letter, initial version:

9th Aug 2021

Dear Dr Gazal,

Thank you for submitting your manuscript entitled "Combining SNP-to-gene linking strategies to pinpoint disease genes and assess disease omnigenicity". As previously noted, we have given the paper our careful consideration and find it of potential interest; but before we send it out to review, we would like the following issues to be addressed.

Firstly, we think that the analysis code must be made fully available for peer review (as it is a fundamental basis of your work). We also require the completed Editorial Policy Checklist and Reporting Summary to ensure your study meets our standards for reproducibility.

We shall hope to receive your revised version as soon as you are able to complete the suggested revisions. If something similar is published in the interim we will have to consider the impact it has on the novelty of a revised manuscript.

If you anticipate a delay of more than four weeks, please let us know. We will be happy to consider your revision so long as nothing similar has been accepted for publication at Nature Genetics or published elsewhere. Should your manuscript be substantially delayed without notifying us in advance and your article is eventually published, the received date may be that of the revised, not the original, version.

If you are not interested in submitting a suitably revised manuscript in the future please let me know immediately so we can close your file. If you have any questions, please contact me.

1) Please ensure that you have completed the Reporting Summary required for review: https://www.nature.com/documents/nr-reporting-summary.pdf

2) Please also complete the Editorial Policy Checklist that would be a requirement for eventual publication in a Nature journal:
https://www.nature.com/documents/nr-editorial-policy-checklist.pdf

Please be aware of our guidelines on digital image standards.

Please use the link below when you are prepared to resubmit.
[REDACTED]

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 www.springernature.com/orcid.

Thank you for your interest in Nature Genetics.

Sincerely,

Michael Fletcher, PhD
Associate Editor, Nature Genetics

ORCID: 0000-0003-1589-7087

Decision Letter, first revision:

13th Oct 2021

Dear Steven,

Your Analysis, "Combining SNP-to-gene linking strategies to pinpoint disease genes and assess disease omnigenicity" has now been seen by 3 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.

Briefly, we think the reports are engaged with your manuscript and sound - broadly - appreciative of the advance presented. However, each referee also raises a number of issues that should be improved

Open Access This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

upon before a potential future publication.

Referee #1 sounds largely positive for your cS2G approach. They provide a list of detailed questions regarding the details of the analysis.
Referee #2 is also positive. Their major concern is how cS2G predictions should be used and interpreted when designing wet lab functional experiments.
Referee #3 is more skeptical. While they provide a short report, we think that the central question - how reliable are these predictions? - is vital and is also mirrored, from a slightly different, wet lab-focused angle, in Reviewer #2's comments.

Overall, we read these reports as guarded support for consideration of a revision; we concluded that a stronger case needs to be made that cS2G does indeed offer the utility that is proposed. In this regard, we think that Referee #2's comments on functional validation is of high priority - ideally the value of cS2G would be demonstrated by some sort of experimental work on a novel prediction, but we acknowledge that this may be non-trivial and we note that this referee also makes a number of specific suggestions for further in silico analysis in this regard. We were also concerned by Referee #1's observation of overlapping training/validation genesets, which may lead to over-estimation of performance, and it would be important to clarify these (and other) details.

To guide the scope of the revisions, the editors discuss the referee reports in detail within the team, including with the chief editor, with a view to identifying key priorities that should be addressed in revision and sometimes overruling referee requests that are deemed beyond the scope of the current study. We hope that you will find the prioritized set of referee points to be useful when revising your study. Please do not hesitate to get in touch if you would like to discuss these issues further.

We therefore invite you to revise your manuscript taking into account all reviewer and editor 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 Analysis format instructions, available
here.
Refer also to any guidelines provided in this letter.

*3) Include a revised version of any required Reporting Summary:

 **Open Access** This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

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 guidelines on digital image standards.

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

[REDACTED]

Note: 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.

Due to the broad scope of the requested revision, we are not placing a hard deadline for submission of the revised manuscript.

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 www.springernature.com/orcid.

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

Sincerely,

Michael Fletcher, PhD
Associate Editor, Nature Genetics

ORCID: 0000-0003-1589-7087

Referee expertise:

Referee #1: statistical genetics, computational genomics.

 **Open Access** This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

Referee #2: genetics, genomics.

Referee #3: statistical genomics.

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
Gazal et al present an approach to combining multiple methods that attempted to link GWAS SNPs to the gene they act through. Their combined approach is found to substantially outperform any individual method, both in terms of accuracy and the proportion of heritability that is linked to the target genes.

The major caveat when assessing the accuracy of approaches to identifying SNP-gene-trait trios is that very few such trios are known, and those trios that have been identified represent a very biased set that is not suitable for use in statistical testing. Instead, Gazal et al approach this problem by considering the heritability coverage in gene-sets selected either from the top 10% of genes with most highly constrained exons and conserved promoters (training) or the top 10% of genes ranked by the PoPS method for a given trait (validation). They also set a heritability baseline of that explained by exonic SNPs, which are considered linked to a gene by definition. While this is not the perfect set of genes to compare against, or a perfect definition of SNPs linked to genes, it represents a serviceable approach. In addition, substantial work is presented that examines the potential biases and inadequacies of the approach taken, demonstrating the results are robust. In fact, every potential major criticism I had of the approach while reading the manuscript was adequately addressed via the extensive robustness analyses.

Specific comments:
The "heritability" metric used throughout this manuscript appears to be always a SNP heritability. However, this is not consistently referred to. e.g. in the definition of "heritability coverage", "precision" and "recall" presented at the start of the results it is presented as "proportion of total disease SNP-heritability", "proportion of disease h2 linked to genes", and "proportion of total disease h2". Also, h2all is used in the formal definition in the methods section, defined as the "heritability explained by common SNPs".

There are several traits with large SNP effects in the MHC region on chromosome 6. Is this region excluded when calculating SNP h2?

The median overlap between the training and validation sets of genes is 20% as opposed to 10% expected by chance. What is the distribution of this overlap? While it is expected there will be overlap in these approaches, this results in the validation sets no longer being independent of the training set. Can the validation sets be filtered to not include any genes in the training set?

The cS2G optimization framework maximised recall while constraining precision to be > 0.75. It is

Open Access This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

unclear whether this constraint was needed in the optimisation approach. It it was, was was the estimate of recall when optimised without constraint?

Pg 11. The extension to include SNP-disease pairs with fine-mapping PIP>0.05 appears to provide little additional insight given the low probability of being causal SNP.

Pg 11. The ~300 diseases examined are defined as being "partially distinct" from the 63 traits used to construct cS2G. Are the results robust to this overlap of diseases?

Pg 13. "computed from N=122K European-ancestry UK Biobank validation samples that were distinct from the N=337K British UK Biobank training samples used for fine-mapping" – were the training and validation sets selected such that individuals in each set were genetically unrelated?

Pg 15. "attained a modest recall of 33%, implying that only 1/3 of causal disease SNPs can be functionally linked to their correct target genes" – The recall is the proportion of heritability correctly attributed to target genes, not the proportion of SNPs. The omnigenicity analysis demonstrated those are distinct measures.

Supp Table 1: with most values being 0, would a better metric be the proportion that are shared non-zero effects? i.e. low correlation due to the non-zero linking scores being poorly correlated, or not overlapping. I note that Table S7 gives that metric for less methods.

Supp Table 3: There are 5 traits with SNP h2 < 0.02 (and a further 2 traits with SNP h2 < 0.1). The estimates of precision and recall from these traits will have very large standard errors. Do these traits contribute equally to the overall estimate of precision/recall.

Supp Table 12: there are lots of "NA" in the gene columns for SNP-Gene-Trait trios – what doe these entries represent. There are also SNP with NA only in the "Confidence Score" column – how are these interpreted?

Supp Table 20: What is the standard error for the proportion of h2(gene) when using all 19995 top genes? The proportion is 1 by definition.

Reviewer #2:
Remarks to the Author:
In this manuscript, Gazal and colleagues explore and combine different strategies to link GWAS SNPs to causal genes. Their main assumption is that a high-quality SNP-to-gene (S2G) strategy should maximize the captured heritability, as assessed using partitioned LD-score methodology. They calibrated different approaches and derived a combined score (cS2G) with good performance, that is with high precision and modest recall as determined using derived gene-sets. At this point, it is important to acknowledge that while the cS2G method performs well in terms of enriching for "causal" genes, it remains largely imprecise to determine if any given gene is causal. This is a limitation of any such methods, but I believe that if should be emphasized more clearly in text, especially as

Open Access This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

investigators might consider using cS2G predictions to design wet-lab experiments (see below).

The Methods section is dense and cryptic to read, but otherwise the methodology appears sound. I did enjoy reading the rest of the text, and thought that the Discussion was quite balanced. The section on the omnigenic model is a nice contribution. I have the following specific comments:

1. cS2G is presented as a powerful approach to help designing functional experiments, but because the work remains based on assumptions and that the predictions are not tested, I don't know if it will have a meaningful impact on experimental designs. The list of cS2G-linked genes is useful for enrichment-type analyses (the omnigenic analyses are a great example). However, a low false positive rate is probably the main motivation for most investigators when designing wet-lab experiments (because resources are limited). When I look at the data (for instance, Figure 3 or Table S13), it would seem like focusing on the Exon strategy, although less "complete", is the most promising strategy when considering the question from a functional lab perspective. Would you agree?

2. The results presented in Figure 4 are interesting, but these predictions are not supported by experimental validations. I would be very interested to know how well the cS2G strategy performs when considering functionally validated SNP-gene pairs? Although it is absolutely true that there are few, a survey of the literature could certainly identify 10-20 such pairs that have been validated (e.g. using CRISPR perturbations). There is the FTO/IRX3/5 locus for BMI. I think that T2D might also have a few; same thing for the blood lipid and blood-cell traits GWAS.

3. Looking at the weights in Table S6, I wonder how much better is the proposed cS2G strategy versus a simpler strategy that would consider only Exon and Promoter variants? I understand that Exon/Promoter link less genes, yet when planning functional experiments (which are rarely comprehensive), higher precision might be more important than recall. Could you quantify this using the "gold-standard" gene-sets?

4. Page 8. "…despite the high weights for Exon and Promoter, 43% of linked common SNPs were not linked to the gene with closest TSS." Provide mean distance between those variants and the linked genes. Do you capture known long-range interactions (eg FTO and IRX3/5 for BMI)?

5. If 43% of cS2G-linked SNPs do not link to the gene with the closest TSS, why taking the closest gene only mildly affects the precision and recall of the approach (Page 8. "This strategy attained only slightly lower precision and recall than the cS2G strategy (0.70 vs. 0.75 and 0.31 vs. 0.33, respectively)")?

6. Make sure that links are provided in the Data Availability section for all datasets used to create S2G strategies. I did not check carefully, but noticed that the Cicero datasets (which are included in the cS2G model) were not listed.

7. In the 2nd paragraph of the Results, you write: "In our primary analyses, each S2G strategy was restricted to the gene(s) with the highest linking score, as we observed that this led to slightly higher precision." I don't understand what this sentence means? Why restricting genes? What is the threshold to define a high linking score?

Open Access This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

8. Somewhere in the manuscript, you should mention why you focus exclusively your analyses to European-ancestry population GWAS results.

9. Did you find examples where two "causal" variants based on high PIP that belong to the same credible set link to different causal genes by the cS2G strategy? If there are such examples, what would be your interpretation?

10. Page 10. "In many instances, multiple causal SNPs were linked to the same gene (e.g. 119 genes were each linked to at least 5 different fine-mapped SNPs for 5 different diseases/traits; Supplementary Table 14), implying that a single gene can be causal for different diseases/traits using different causal SNP-gene links." Does that imply that the different "causal" SNPs are not in LD with each other?

11. Figure 4. Are the p-values on the y-axis GWAS p-values?

13. Page 12. "In summary, our cS2G strategy validated 205 previously curated disease-associated SNP-gene pairs…". I don't think that these predictions are validated but rather that they are retrieved consistently across 2 different methods.

Signed: Guillaume Lettre

Reviewer #3:
Remarks to the Author:
In this paper, Gazal and colleagues analyze 50 SNP-to-gene (S2G) linking methods and attempt to define optimal combinations for identifying disease genes by assessing the heritability explained by using a certain S2G strategy. They use GWAS data on 63 traits to come up with an optimal combination strategy (cS2G) that they subsequently apply to fine-mapping results of 49 UK Biobank diseases and identify 7,111 SNP-gene-disease triplets. It is clear that the authors have put a substantial amount of effort into the study, and the problem of how to best link SNPs to causal genes is of course important. However, I think the optimal S2G strategy will vary from one locus to another and am not convinced that the proposed combined strategy is necessarily the best. It is also unclear whether the proposed combination strategy is optimal as there are many adhoc choices along the way, and it is difficult to assess the effect of such choices on performance.

I have two major comments, as follows:

A. The authors state that the different linking methods have low correlation among themselves, so it makes sense to combine them. But how to best combine them is not trivial given the many differences among the methods and the different scenarios that can occur in a GWAS setting. Certain methods perform well under certain scenarios, e.g. depending on where the variant may be located, knowledge of relevant tissue-/cell-type, steady state vs. dynamic eQTLs etc. The authors propose a linear combination with weights optimized to improve precision/recall. But is that an optimal approach for a

Open Access This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

given locus? The authors compare with individual strategies, but each individual strategy might perform well under particular scenarios.
B. I find it difficult to interpret the results on the UK Biobank. How can we know those 7,111 SNP-gene-disease triplets are "high-confidence" results? Fine-mapping at any single GWAS locus is generally challenging and there are many caveats to such results, so how can we confidently identify so many disease genes in one analysis?
3. Similar comments about the results on checking the omnigenic model. Those are quite speculative in nature, and difficult to verify.

Minor comment:
The paper is very densely written, which interferes with the flow of the paper.

Author Rebuttal, first revision:

Response to reviewers for NG-AN57945R1 (Gazal et al.)

Reviewer #1:
Remarks to the Author:
Gazal et al present an approach to combining multiple methods that attempted to link GWAS SNPs to the gene they act through. Their combined approach is found to substantially outperform any individual method, both in terms of accuracy and the proportion of heritability that is linked to the target genes.

The major caveat when assessing the accuracy of approaches to identifying SNP-gene-trait trios is that very few such trios are known, and those trios that have been identified represent a very biased set that is not suitable for use in statistical testing. Instead, Gazal et al approach this problem by considering the heritability coverage in gene-sets selected either from the top 10% of genes with most highly constrained exons and conserved promoters (training) or the top 10% of genes ranked by the PoPS method for a given trait (validation). They also set a heritability baseline of that explained by exonic SNPs, which are considered linked to a gene by definition. While this is not the perfect set of genes to compare against, or a perfect definition of SNPs linked to genes, it represents a serviceable approach. In addition, substantial work is presented that examines the potential biases and inadequacies of the approach taken, demonstrating the results are robust. In fact, every potential major criticism I had of the approach while reading the manuscript was adequately addressed via the extensive robustness analyses.

We thank the reviewer for the accurate summary of our work, and for suggesting that our results are robust.

Specific comments:

Open Access This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

1. The “heritability” metric used throughout this manuscript appears to be always a SNP heritability. However, this is not consistently referred to. e.g. in the definition of “heritability coverage”, “precision” and “recall” presented at the start of the results it is presented as “proportion of total disease SNP-heritability”, “proportion of disease h2 linked to genes”, and “proportion of total disease h2”. Also, h2all is used in the formal definition in the methods section, defined as the “heritability explained by common SNPs”.

The reviewer is correct that our metrics are based on SNP-heritability. We agree that it is our responsibility to provide a clear exposition. We have updated the Abstract (p. 2), Introduction (p. 3), *Overview of methods* subsection of the Results section (p. 4), and *Leveraging the combined S2G strategy to empirically assess disease omnigenicity* subsection of the Results section (p. 16) to specifically refer to SNP-heritability. We still use the notation h^2, but are now very clear when introducing this notation that it refers to SNP-heritability (*Overview of methods* subsection of the Results section, p. 4). (We prefer to avoid other notation such as h_g^2, which would lead to the awkward notation $h_{g^2,gene}$, but are open to using such notation instead if reviewers and/or editors have a strong preference.) We have also modified h^2_{all} to h^2 in the Methods section (p. 32).

2. There are several traits with large SNP effects in the MHC region on chromosome 6. Is this region excluded when calculating SNP h2?

The reviewer is correct that the MHC region contains large SNP effects for some traits. In all S-LDSC analyses, we removed from the GWAS summary statistics all SNPs in the MHC region in order to estimate S-LDSC per-SNP heritability parameters, but we kept all SNPs in the MHC when predicting per-SNP heritability using functional annotations (as recommended in Finucane et al. 2015 Nat Genet; ref. 11). Similarly, all SNPs in the MHC region were removed from the fine-mapping analyses and the gene ranking analyses (analogous to Weissbrod et al. 2020 Nat Genet; ref. 7). We have updated the Methods section (p. 32, 36 and 37), citing ref. 11 and ref. 7, to clarify this point.

3. The median overlap between the training and validation sets of genes is 20% as opposed to 10% expected by chance. What is the distribution of this overlap? While it is expected there will be overlap in these approaches, this results in the validation sets no longer being independent of the training set. Can the validation sets be filtered to not include any genes in the training set?

The reviewer has raised 2 related questions: (i) what is the distribution of the overlap between the training and validation sets of genes, across the 63 diseases/traits analyzed; and (ii) can the analyses be repeated while filtering the validation sets to not include any genes in the training set. We address each question in turn.

Open Access This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

(i) what is the distribution of the overlap between the training and validation sets of genes, across the 63 diseases/traits analyzed.

Across the 63 diseases/traits analyzed, the % overlap between the training gene set (which does not vary across disease/traits) and the validation gene sets (which does vary across diseases/traits) had a median of 20%, mean of 20%, standard deviation of 4.1%, and range from 13%-28%. The distribution is described in detail in our new Supplementary Figure 2. We have updated the *Overview of methods* subsection of the Results section (p. 5) to cite the Methods section and Supplementary Figure 2, and we have updated the Methods section (p. 34) to include these quantifications and cite Supplementary Figure 2.

(ii) can the analyses be repeated while filtering the validation sets to not include any genes in the training set.

We agree that this is a valuable check. We have repeated our analyses while removing all 1,760 genes in the training gene set from the set of all genes analyzed in the validation step, ensuring that the 1,760 genes do not appear in the validation gene sets (and that the validation gene sets have no enrichment or depletion of genes in the training gene set). We estimated a precision of 0.80 (s.e. = 0.09) for cS2G, which is similar to (and actually slightly higher than) what we reported in our primary analysis (precision of 0.75, s.e. = 0.06). We have updated the *Combining S2G strategies* subsection of the Results section (p.10) to note this result.

4. The cS2G optimization framework maximised recall while constraining precision to be > 0.75. It is unclear whether this constraint was needed in the optimisation approach. What was the estimate of recall when optimised without constraint?

The reviewer has raised 2 related questions: (i) why did we constrain precision to be ≥0.75, and (ii) how do our estimates of recall change when this constraint is removed? We address each question in turn.

(i) why did we constrain precision to be ≥0.75?

Our goal was to construct a combined S2G strategy that maximizes the recall while providing high precision to maximize the utility of functional follow-up studies. (Indeed, Reviewer #2 has suggested that “a low false positive rate is probably the main motivation for most investigators when designing wet-lab experiments”; see response to Reviewer #2 Comment 1 below.) We have updated the *Overview of methods* subsection of the Results section (p. 5) and the Methods section (p. 34) to clarify this point.

(ii) how do our estimates of recall change when this constraint is removed?


Open Access This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

We first consider our primary analysis. The primary cS2G strategy attained a precision of 0.75 (s.e. 0.06) and recall of 0.33 (s.e. 0.03), meta-analyzed across diseases/traits. The cS2G strategy (including its weights) and resulting precision and recall were unchanged when removing the constraint that precision must be ≥0.75; however, we still recommend constraining precision to be ≥0.75, to maximize the utility of functional follow-up studies. We have updated the *Combining S2G strategies* subsection of the Results section (p. 10) to include this result.

We next consider our secondary analyses. The removal of the constraint that precision must be ≥0.75 is pertinent only to the secondary analyses numbered as "Third", "Fourth", and "Sixth" in the *Combining S2G strategies* subsection of the Results section (p. 11). For the secondary analyses numbered as "Fourth" and "Sixth", the cS2G strategy that we constructed was unchanged when removing the constraint that precision must be ≥0.75. For the secondary analysis numbered as "Third" (in which we added the 3 non-functionally informed main S2G strategies from Table 1; Gene body, Gene±100kb and Closest TSS), the precision of 0.75 (s.e. 0.06) and recall of 0.33 (s.e. 0.03) (identical to our primary analysis) changed to a precision of 0.34 (s.e. 0.03) and recall of 0.34 (s.e. 0.03) when removing the constraint that precision must be ≥0.75. We consider this to be an unfavorable result, demonstrating the advantages of constraining the precision to be ≥0.75. We have updated the *Combining S2G strategies* subsection of the Results section (p. 11) to include these results.

5. Pg 11. The extension to include SNP-disease pairs with fine-mapping PIP>0.05 appears to provide little additional insight given the low probability of being causal SNP.

We agree that the analysis of SNP-disease pairs with PIP>0.05 (instead of PIP>0.5) is of low interest -particularly as we now prioritize SNP-gene-disease triplets with confidence score >0.5, which is not possible when PIP<0.5 (see response to Reviewer #2 Comment 1). We have removed this analysis from the *Leveraging the combined S2G strategy to pinpoint disease genes* subsection of the Results section (p. 15), and updated the Methods section (p. 37) to note that results for SNP-disease pairs with PIP>0.05 are provided in Data Availability, for completeness.

For the same reason, we have removed the analysis of SNP-disease pairs from the NHGRI GWAS catalog (which have low probability of being causal SNPs) from the *Leveraging the combined S2G strategy to pinpoint disease genes* subsection of the Results section (p. 15), and updated the Methods section (p. 37) to note that results for SNP-disease pairs from the NHGRI GWAS catalog are provided in Data Availability. We are open to restoring this analysis to the main text if reviewers and/or editors have a strong preference.

6. Pg 11. The ~300 diseases examined are defined as being "partially distinct" from the 63 traits used to construct cS2G. Are the results robust to this overlap of diseases?

Open Access This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

The reviewer raises a valid point that the 63 diseases/traits used to construct cS2G may overlap the ~300 diseases/traits included in the Open Targets curated list of 577 linked sentinel SNP-gene pairs.

To address this, we manually removed 247 of the 577 Open Targets SNP-gene pairs by removing all pairs for the Open Targets diseases/traits with a name identical or similar to the names of our 63 diseases/traits. We repeated the Open Targets analysis while restricting to the remaining 330 SNP-gene pairs. We determined that results were little changed: 211 of the 330 SNPs had a linked gene with cS2G linking score >0.5, and the cS2G prediction of the target gene matched the Open Targets prediction for 120 of these 211 SNPs (57%, vs. 205/368=58% when considering the complete Open Targets curated list of 577 SNP-gene pairs).

We have now added an analysis of 1,668 linked sentinel SNP-gene pairs (for ~100 diseases and complex traits) from Weeks et al. medrxiv (ref. 51) (see response to Reviewer #2 Comment 2); this analysis is analogous to the analysis of 577 linked sentinel SNP-gene pairs from Open Targets. Once again, analogously, the 63 diseases/traits used to construct cS2G may overlap the ~100 diseases/traits included in the Weeks et al. curated list of 1,668 linked sentinel SNP-gene pairs. Thus, we repeated the above robustness analysis, manually removing 696 of the 1,668 Weeks et al. SNP-gene pairs by removing all pairs for the Weeks et al. diseases/traits with a name identical or similar to the names of our 63 diseases/traits. We repeated the Weeks et al. analysis while restricting to the remaining 929 SNP-gene pairs. We determined that results were little changed: 677 of the 929 SNPs had a linked gene with cS2G linking score >0.5, and the cS2G prediction of the target gene matched the Weeks et al. prediction for 462 of these 677 SNPs (68%, vs. 741/1124=66% when considering the complete Weeks et al. curated list of 1,668 SNP-gene pairs).

We note that although our manual curation step to remove SNP-gene pairs for Open Targets diseases/traits (resp. Weeks et al. diseases/traits) that overlap our 63 diseases/traits is likely to be incomplete, and some overlapping or genetically correlated traits may remain, the fact that results were not sensitive to a substantial removal of overlapping diseases/traits strongly supports the overall robustness of our results.

We have included this information in the new *Validating the combined S2G strategy using curated lists of disease-associated SNP-gene pairs* subsection of the Results section (p. 12-13), citing a new Supplementary Table 15.

7. Pg 13. “computed from N=122K European-ancestry UK Biobank validation samples that were distinct from the N=337K British UK Biobank training samples used for fine-mapping” – were the training and validation sets selected such that individuals in each set were genetically unrelated?

Open Access This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

The N=122K European-ancestry UK Biobank validation samples consisted of all N=459K European-ancestry UK Biobank samples minus the N=337K unrelated British UK Biobank samples that were used for training (analogous to Weissbrod et al. 2020 Nat Genet; ref. 7). Thus, some of the validation samples may be related to training samples. We agree that it is important to check that this does not impact our results.

We thus repeated our analyses using a new set of N=49K non-British European-ancestry UK Biobank samples that were chosen to be unrelated to the N=337K British UK Biobank samples that were used for training. Using this set of validation samples, we determined that the top 200 (resp. top 2,000) genes explained 52±7% (resp. 99±9%) of the disease heritability linked to genes in cis using the cS2G strategy (h^2_{gene}, which captures 55±4% of h^2. These results were very similar to our results using the N=122K validation samples, in which we determined that the top 200 (resp. top 2,000) genes explained 52±6% (resp. 96±8%) of the disease heritability linked to genes in cis using the cS2G strategy (h^2_{gene}, which captures 53±3% of h^2).

We have updated the *Leveraging the combined S2G strategy to empirically assess disease omnigenicity* subsection of the Results section (p. 16), citing a new Supplementary Figure 10 to include this result.

8. Pg 15. “attained a modest recall of 33%, implying that only 1/3 of causal disease SNPs can be functionally linked to their correct target genes” – The recall is the proportion of heritability correctly attributed to target genes, not the proportion of SNPs. The omnigenicity analysis demonstrated those are distinct measures.

We agree with the reviewers that these two quantities are distinct. We have modified this text to state that only 1/3 of disease SNP-heritability can be explained by causal disease SNPs functionally linked to their correct target genes (Discussion section, p. 18).

9. Supp Table 1: with most values being 0, would a better metric be the proportion that are shared non-zero effects? i.e. low correlation due to the non-zero linking scores being poorly correlated, or not overlapping. I note that Table S7 gives that metric for less methods.

We agree with the reviewer that the proportion that are shared non-zero effects is an informative metric. We now include both correlation values and overlap proportion values in the “Overview of Methods” subsection of the Results section (p. 4), and define overlap proportion in the Method section (p. 31). We added these new overlap values in a Supplementary Table 1a, and retained the correlations as Supplementary Table 1b.

10. Supp Table 3: There are 5 traits with SNP h2 < 0.02 (and a further 2 traits with SNP h2 < 0.1). The estimates of precision and recall from these traits will have very large standard errors. Do these traits contribute equally to the overall estimate of precision/recall.


Open Access This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

The reviewer is correct that traits with low SNP-heritability can generate large standard errors. Nevertheless, we included 5 traits with SNP-heritability < 0.02 because they had z-score>6 for non-zero SNP-heritability, following the recommendation of Gazal et al. 2017 Nat Genet (ref. 52).

However, we performed fixed effect meta-analyses when computing overall estimates of precision/recall, thus giving low weights to traits with large standard errors. Thus, inclusion of the 5 traits with SNP-heritability < 0.02 had little impact on our results. (For example, when removing the 5 traits with SNP-heritability < 0.02, estimates of precision changed from 0.747 (s.e. 0.061) to 0.753 (s.e. 0.062) and estimates of recall changed from 0.330 (s.e. 0.027) to 0.332 (s.e. 0.027)).

We have updated the Methods section (p. 33) to clarify these points.

11. Supp Table 12: there are lots of "NA" in the gene columns for SNP-Gene-Trait trios – what do these entries represent. There are also SNP with NA only in the "Confidence Score" column – how are these interpreted?

We apologize for this confusion. Fine-mapped SNPs that are not linked to genes by the cS2G strategy have a value of NA for the columns "Gene", "cS2G score*" (the cS2G score before normalization), "cS2G score" (the cS2G score after normalization), and "cS2G confidence score". In addition, fine-mapped SNPs with cS2G score <0.5 have a value of NA for the column "cS2G confidence score", as we do not report causal SNP-gene-disease triplets for such SNPs in this Table.

We have updated the caption of Supplementary Table 17 (formerly Supplementary Table 12) to clarify these points.

12. Supp Table 20: What is the standard error for the proportion of h2(gene) when using all 19995 top genes? The proportion is 1 by definition.

The proportion of h^2_{gene} explained by the top X genes was computed as the SNP-heritability explained by the top X genes divided by h^2_{gene}. The standard error of the proportion of h^2_{gene} explained by the top X genes was computed as the standard error of the SNP-heritability explained by the top X genes divided by the point estimate of h^2_{gene}. Computing the standard error of a ratio is challenging, but we believe this to be a reasonable approximation, as the numerator has greater uncertainty (i.e. standard error / point estimate) than the denominator when X<19,995, and the errors are correlated such that this approximation is conservative. (We note that the numerator and denominator are each meta-analyzed across diseases/traits; a ratio of meta-analyzed values is more robust than a meta-analysis of ratios.)


Open Access This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

In the special case where X=19,995 (all genes), we agree that the proportion is equal to 1 by definition, and its standard error is equal to 0 (instead of the value greater than 0 resulting from the conservative approximation). We have thus updated Figure 5a and Supplementary Table 23 (formerly Supplementary Table 20) to set the standard error to 0 in this case. For cases where X<19,995, we note that values greater than 1 are outside the biologically plausible 0-1 range, but allowing point estimates outside the biologically plausible 0-1 range is necessary to ensure unbiasedness.

We have updated the Methods section (p. 38) and the captions of Figure 5a, Supplementary Table 22 and Supplementary Table 23 (formerly Supplementary Table 19 and Supplementary Table 20) to clarify these points.

Open Access This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

Reviewer #2:
Remarks to the Author:
In this manuscript, Gazal and colleagues explore and combine different strategies to link GWAS SNPs to causal genes. Their main assumption is that a high-quality SNP-to-gene (S2G) strategy should maximize the captured heritability, as assessed using partitioned LD-score methodology. They calibrated different approaches and derived a combined score (cS2G) with good performance, that is with high precision and modest recall as determined using derived gene-sets. At this point, it is important to acknowledge that while the cS2G method performs well in terms of enriching for "causal" genes, it remains largely imprecise to determine if any given gene is causal. This is a limitation of any such methods, but I believe that if should be emphasized more clearly in text, especially as investigators might consider using cS2G predictions to design wet-lab experiments (see below).

The Methods section is dense and cryptic to read, but otherwise the methodology appears sound. I did enjoy reading the rest of the text, and thought that the Discussion was quite balanced. The section on the omnigenic model is a nice contribution.

We thank the reviewer for the accurate summary of our work, and for suggesting that the methodology is sound and our work is a nice contribution. We agree with the limitation noted by the reviewer (see response to Reviewer #2 Comments 1 and 2).

I have the following specific comments:

1. cS2G is presented as a powerful approach to help designing functional experiments, but because the work remains based on assumptions and that the predictions are not tested, I don't know if it will have a meaningful impact on experimental designs. The list of cS2G-linked genes is useful for enrichment-type analyses (the omnigenic analyses are a great example). However, a low false positive rate is probably the main motivation for most investigators when designing wet-lab experiments (because resources are limited). When I look at the data (for instance, Figure 3 or Table S13), it would seem like focusing on the Exon strategy, although less "complete", is the most promising strategy when considering the question from a functional lab perspective. Would you agree?

The reviewer makes a good point that low false positive rate (i.e. precision) is of paramount importance to maximizing the utility of functional follow-up studies (also see part (i) of response to Reviewer #1 Comment 4). This raises the question of whether to prioritize S2G strategies with higher precision (but lower recall), e.g. Exon as compared to cS2G.

In fine-mapping analyses of UK Biobank traits, we previously focused on the estimated number of correct and incorrect SNP-gene-disease triplets (Figure 3; now Figure 3a). However, motivated by the reviewer's comment, we have now elected to focus on the confidence score of

Open Access This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

each SNP-gene-disease triplet; the confidence score is defined as the product of the PIP of the constituent fine-mapped SNP and the estimated precision of the SNP-gene-disease triplet. In particular, we advocate for prioritizing triplets with confidence score > 0.50, to maximize the utility of functional follow-up studies. We have added a new main Figure 3b reporting the distribution of confidence scores of SNP-gene-disease triplets for each S2G strategy. We have also computed the number of triplets with confidence score > 0.50, which is equal to 5,095 for cS2G (vs. 2,763 for Exon). We thus believe that cS2G substantially outperforms the Exon strategy towards the goal of identifying high-confidence triplets for functional follow-up studies.

We have updated the *Leveraging the combined S2G strategy to pinpoint disease genes* subsection of the Results section (p. 14), citing the new Figure 3b and an updated Supplementary Table 18 (formerly Supplementary Table 13), to note these results. We have also modified the manuscript to place greater emphasis on the confidence score of each triplet, specifically emphasizing the 5,095 SNP-gene-disease triplets with confidence score > 0.50. We have modified the Abstract (p. 2), *Leveraging the combined S2G strategy to pinpoint disease genes* subsection of the Results section (p. 14), and Discussion section (p. 18) accordingly.

2. The results presented in Figure 4 are interesting, but these predictions are not supported by experimental validations. I would be very interested to know how well the cS2G strategy performs when considering functionally validated SNP-gene pairs? Although it is absolutely true that there are few, a survey of the literature could certainly identify 10-20 such pairs that have been validated (e.g. using CRISPR perturbations). There is the FTO/IRX3/5 locus for BMI. I think that T2D might also have a few; same thing for the blood lipid and blood-cell traits GWAS.

We agree with the reviewer that it is of high interest to assess the cS2G strategy using experimentally validated SNP-gene pairs (even though few such pairs exist). We have now identified 17 disease/trait loci (including 12 loci reported in the review paper of Gallagher et al. 2018 Am J Hum Genet; ref. 58) containing 25 experimentally validated causal SNP-gene pairs. These include the rs1421085-*IRX5/IRX3* pair for BMI (not strictly a SNP-gene pair, as rs1421085 is linked to both genes), the rs11257655-*CAMK1D* pair for T2D, the rs12740374-*SORT1* pair for LDL and the rs737092-*RBM38* pair for red blood cell count. We note that we did not include examples where the gene has been functionally validated but the causal variant has not been identified, and we did not include examples validated only using functional data (such as histone modifications or gene expression) or overlapping TF motifs.

Of the 25 experimentally validated causal SNP-gene pairs, 16 had the causal SNP annotated with cS2G linking score >0.5, of which 11 had the causal SNP accurately linked to the validated causal gene. We thus estimated precision as 11/16 = 0.69 (s.e. = 0.12), and recall as 11/25 = 0.44 (s.e. = 0.10). This is a lower precision and higher recall than our estimates based on validation critical gene sets (0.75 for precision, 0.33 for recall), but the differences were not statistically significant due to the small number of experimentally validated SNP-gene pairs. We note that cS2G obtained the 2nd best precision and the best recall when compared to its

Open Access This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

constituent strategies (max precision = 0.78, s.e. = 0.14 for EpiMap), and higher precision than the Closest TSS strategy (0.56, s.e. = 0.10), but differences were not significant due to the noise of the N=25 sample size. We note that the *FTO*/*IRX3*/*IRX5* locus mentioned by the reviewer was included in these analyses, but the causal SNP rs1421085 was not annotated with cS2G linking score >0.5, as it was linked to 2 genes (*FTO* and *RPGRIP1L*) each with a score of 0.5, and we decided (prior to these analyses; see originally submitted manuscript) to prioritize SNPs linked to a gene with cS2G linking score >0.5 (if we had instead used cS2G linking score ≥0.5 as our threshold, the impact on results would be minimal, with precision = 11/17 = 0.65 instead of 0.69). Overall, we consider this to be a promising validation of the potential of cS2G to identify causal SNP-gene pairs. We have included these new results in a new *Validating the combined S2G strategy using curated lists of disease-associated SNP-gene pairs* subsection of the Results section (p. 12), citing a new Table 2 and new Supplementary Tables 12 and 13. We have also moved the large amount of text interpreting Figure 4a-d to a new Supplementary Note (see response to Reviewer #3 Comment 4), while retaining Figure 4 as a main Figure.

To further support our cS2G predictions, we previously analyzed a curated disease-associated list of 577 linked sentinel SNP-gene pairs with the underlying genes validated with high confidence by Open Targets (now published in Mountjoy et al. 2021 Nat Genet; ref. 42), with the caveat that these do not represent experimentally validated SNP-gene pairs. We obtained a precision of 0.58 and recall of 0.36. We note that these results are impacted by the fact that Open Targets reports sentinel SNPs rather than causal fine-mapped SNPs (see Supplementary Table 16, formerly Supplementary Table 17, for verification of this statement using UK Biobank fine-mapping analyses). These results are now reported in the new *Validating the combined S2G strategy using curated lists of disease-associated SNP-gene pairs* subsection of the Results section (p. 12).

We have now analyzed a separate curated disease-associated list of 1,668 sentinel SNP-gene pairs validated using nearby fine-mapped protein-coding SNPs (Weeks et al. medrxiv; ref 51), again with the caveat that these do not represent experimentally validated SNP-gene pairs. We observed a precision of 0.66 and recall of 0.46. We note that these results are impacted by the fact that the Weeks et al. curated list assumes that a gene with a protein-coding fine-mapped SNP is always targeted by 1Mb surrounding non-coding fine-mapped SNPs. We have updated the new *Validating the combined S2G strategy using curated lists of disease-associated SNP-gene pairs* subsection of the Results section to note these results (p. 12).

3. Looking at the weights in Table S6, I wonder how much better is the proposed cS2G strategy versus a simpler strategy that would consider only Exon and Promoter variants? I understand that Exon/Promoter link less genes, yet when planning functional experiments (which are rarely comprehensive), higher precision might be more important than recall. Could you quantify this using the “gold-standard” gene-sets?


Open Access This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

We agree that it is of interest to consider an S2G strategy consisting of the both Exon and Promoter links. We have now analyzed this strategy, and determined that it achieved a precision of 1.05 (s.e. 0.07) (statistically indistinguishable from the Exon (precision of 1 by definition) and Promoter (0.80, s.e. = 0.16) strategies) and recall of 0.16 (s.e. 0.01) (statistically indistinguishable from the sum of recalls attained by Exon (0.10) and Promoter (0.05) strategies), which is much lower than the recall of 0.33 (s.e. 0.03) attained by the cS2G strategy. We have updated the *Combining S2G strategies* subsection of the Results section (p. 10, citing updated Supplementary Table 5), to include results for the Exon + Promoter strategy.

In fine-mapping analyses of UK Biobank traits, we have now also computed the number of triplets with confidence score > 0.50 (see response to Reviewer #2 Comment 1) for the Exon + Promoter strategy. There were 3,530 such triplets, substantially lower than the 5,095 such triplets identified using the cS2G strategy. We have updated Supplementary Table 18 (formerly Supplementary Table 13) to include this result.

4. Page 8. "…despite the high weights for Exon and Promoter, 43% of linked common SNPs were not linked to the gene with closest TSS." Provide mean distance between those variants and the linked genes. Do you capture known long-range interactions (eg FTO and IRX3/5 for BMI)?

We thank the reviewer for this suggestion. The mean distance to the gene TSS for all cS2G links involving common SNPs was 96kb (mean of 24kb for the 57% of linked SNPs linked to the gene with closest TSS, mean of 192kb for the 43% of linked SNPs not linked to the gene with closest TSS). We have updated the *Combining S2G strategies* subsection of the Results section (p. 9, citing a new Supplementary Figure 8), to note these results.

We agree that evaluation of cS2G using experimentally validated SNP-gene-disease triplets is of high interest. We provide a description of these new analyses in our response to Reviewer #2 Comment 2.

5. If 43% of cS2G-linked SNPs do not link to the gene with the closest TSS, why taking the closest gene only mildly affects the precision and recall of the approach (Page 8. "This strategy attained only slightly lower precision and recall than the cS2G strategy (0.70 vs. 0.75 and 0.31 vs. 0.33, respectively)")?

The reviewer is correct that 43% of cS2G-linked SNPs do not link to the gene with closest TSS, and the reviewer is also correct that the strategy of taking cS2G-linked SNPs and linking them to the closest gene (ClosestTSS-cS2GSNPs) attained only slightly lower precision and recall than cS2G (0.70 vs. 0.75 and 0.31 vs. 0.33, respectively; we note that the slightly lower recall despite same set of SNPs is a consequence of the slightly lower precision). Indeed, it is possible for two cS2G strategies to have very different SNP-gene links but attain relatively

Open Access This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

similar performance, if both strategies have imperfect SNP-gene links. We have updated the *Combining S2G strategies* subsection of the Results section (p. 9-10) to clarify this point.

We sought to further investigate the reduction in precision of ClosestTSS-cS2GSNPs vs. cS2G. We estimated the reduction in precision of ClosestTSS-cS2GSNPs as "proportion of h2 coverage explained by cS2G SNPs not linked to Closest TSS by cS2G" * ("precision of cS2G for cS2G SNPs not linked to Closest TSS by cS2G" – "precision of Closest TSS for cS2G SNPs not linked to Closest TSS by cS2G") = 0.33 * (0.64 – 0.55) = 0.03, consistent with the observed reduction of 0.05. We have updated the caption of Supplementary Table 8 to include this computation.

6. Make sure that links are provided in the Data Availability section for all datasets used to create S2G strategies. I did not check carefully, but noticed that the Cicero datasets (which are included in the cS2G model) were not listed.

We apologize for this mistake. We have added a new Supplementary Table 26 providing links for all datasets used to create S2G strategies, and we now cite this Table in the Data Availability section (p. 21). We also now provide at https://alkesgroup.broadinstitute.org/cS2G/S2G_datasets/ all the datasets that were obtained through personal communication (Cicero), downloaded from webpages (Ensembl list of TSS), or through fine-mapping (GTeX and eQTLGen fine-mapped cis-eQTLs). We have updated the Data Availability section accordingly.

7. In the 2nd paragraph of the Results, you write: “In our primary analyses, each S2G strategy was restricted to the gene(s) with the highest linking score, as we observed that this led to slightly higher precision.” I don’t understand what this sentence means? Why restricting genes? What is the threshold to define a high linking score?

Each S2G strategy came with varying definitions of raw linking values. For example, for the S2G strategies based on *cis*-eQTL, the raw linking value is defined as $-\log_{10}$ of the association P value; for the S2G strategies based on fine-mapped eQTL, the raw linking value is defined as the causal posterior probability (CPP). For most S2G strategies, we did not impose a threshold on the raw linking value. (The only exception is the fine-mapped eQTL S2G strategies, for which we imposed a threshold of CPP ≥0.05.) Thus, S2G strategies include instances of SNP-gene links with low linking scores. We have updated the *Overview of methods* subsection of the Results section (p. 4) and the Methods section (p. 29 and 31) to clarify this point.

We restricted each S2G strategy such that each SNP was restricted to the gene(s) with the highest linking score (regardless of whether this linking score was high or low in absolute terms; no specific threshold). We made this choice because we observed that this led to higher precision than retaining all SNP-gene links reported by the raw S2G strategy (see Supplementary Figure 1, formerly Supplementary Figure 12). We have updated the *Overview of*

Open Access This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

methods subsection of the Results section (p. 4 citing Supplementary Figure 1) to clarify this point. (We have retained the content in the Discussion section (p. 19) noting that this choice does not reflect biological reality, in which a regulatory element may target more than one gene, and noting that refinements to this choice are a direction for future research.)

In our application to link UK Biobank (candidate) causal variants to their target genes, we restricted to SNPs that had a linked gene with cS2G linking score >0.5, consistent with the goal of providing a single gene with high precision to maximize the utility of functional follow-up studies (see response to Reviewer #2 Comment 1). We note that we have now determined that restricting cS2G to those links leads to similar results in our polygenic analyses of precision and recall; Supplementary Table 8. We have updated the new *Validating the combined S2G strategy using curated lists of disease-associated SNP-gene pairs* subsection of the Results section (p. 12) and Supplementary Table 8 to clarify this point.

8. Somewhere in the manuscript, you should mention why you focus exclusively your analyses to European-ancestry population GWAS results.

We thank the reviewer for this suggestion. We focused on European-ancestry GWAS results for two reasons. First, they are currently available in far greater sample sizes than GWAS from other populations. Second, the functionally informed S2G strategies that we analyze are primarily based on functional experiments in European-ancestry samples. We agree that assessing transferability to non-European populations is a critical future research direction. We have updated the Discussion section (p. 20) to clarify these points.

9. Did you find examples where two “causal” variants based on high PIP that belong to the same credible set link to different causal genes by the cS2G strategy? If there are such examples, what would be your interpretation?

We thank the reviewer for this suggestion. We agree that it is appropriate to assess how often two high-PIP SNPs for the same disease/trait at the same locus were linked by cS2G to the same gene vs. different genes. Among the 7,111 causal SNP-gene-disease triplets involving PIP>0.5 SNPs (for UK Biobank diseases/traits) that had a linked gene with cS2G linking score >0.5, we identified 700 pairs of PIP>0.5 SNPs for the same disease/trait within 10kb of each other. Of these 700 pairs, 555 (79%) were linked to the same gene, whereas 145 (21%) were linked to different genes. Assuming that these pairs truly target the same gene, it is not surprising that 79%<100%, as the SNP-gene-disease triplets are expected to be imperfect (mean confidence score = 0.77 across the 700 pairs) due to imperfect fine-mapping (PIP<1) and/or imperfect SNP-gene linking (cS2G linking score <1). Indeed, SNP-gene-disease triplets for the 555 concordant pairs (linked to the same gene) had a mean confidence score of 0.79, vs. 0.68 for the 145 discordant pairs.

 **Open Access** This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

We have updated the *Leveraging the combined S2G strategy to pinpoint disease genes* subsection of the Results section (p. 14), to note these results.

10. Page 10. “In many instances, multiple causal SNPs were linked to the same gene (e.g. 119 genes were each linked to at least 5 different fine-mapped SNPs for 5 different diseases/traits; Supplementary Table 14), implying that a single gene can be causal for different diseases/traits using different causal SNP-gene links.” Does that imply that the different “causal” SNPs are not in LD with each other?

The reviewer is correct that these different “causal” SNPs tend to be not in LD. We identified 3,900 pairs of fine-mapped SNPs for different diseases/traits that were linked to the same gene. These 3,900 pairs of SNPs had mean r^2 = 0.09. We have updated the *Leveraging the combined S2G strategy to pinpoint disease genes* subsection of the Results section (p. 14) to note this result.

11. Figure 4. Are the p-values on the y-axis GWAS p-values?

We apologize for this confusion. The p-values on the y-axis are indeed GWAS p-values. We have updated the caption of Figure 4 to clarify this point.

13. Page 12. “In summary, our cS2G strategy validated 205 previously curated disease-associated SNP-gene pairs…”. I don’t think that these predictions are validated but rather that they are retrieved consistently across 2 different methods.

The reviewer is correct that these predictions are not validated but rather are consistent with Open Targets results. We have removed this text; this analysis has now been moved to the new *Validating the combined S2G strategy using curated lists of disease-associated SNP-gene pairs* subsection of the Results section (p. 12).

Signed: Guillaume Lettre

Open Access This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

Reviewer #3:
Remarks to the Author:
In this paper, Gazal and colleagues analyze 50 SNP-to-gene (S2G) linking methods and attempt to define optimal combinations for identifying disease genes by assessing the heritability explained by using a certain S2G strategy. They use GWAS data on 63 traits to come up with an optimal combination strategy (cS2G) that they subsequently apply to fine-mapping results of 49 UK Biobank diseases and identify 7,111 SNP-gene-disease triplets. It is clear that the authors have put a substantial amount of effort into the study, and the problem of how to best link SNPs to causal genes is of course important. However, I think the optimal S2G strategy will vary from one locus to another and am not convinced that the proposed combined strategy is necessarily the best. It is also unclear whether the proposed combination strategy is optimal as there are many adhoc choices along the way, and it is difficult to assess the effect of such choices on performance.

We thank the reviewer for the accurate summary of our work, and for suggesting that the problem is important. We respond to the reviewer concerns below.

I have two major comments, as follows:

1. The authors state that the different linking methods have low correlation among themselves, so it makes sense to combine them. But how to best combine them is not trivial given the many differences among the methods and the different scenarios that can occur in a GWAS setting. Certain methods perform well under certain scenarios, e.g. depending on where the variant may be located, knowledge of relevant tissue-/cell-type, steady state vs. dynamic eQTLs etc. The authors propose a linear combination with weights optimized to improve precision/recall. But is that an optimal approach for a given locus? The authors compare with individual strategies, but each individual strategy might perform well under particular scenarios.

We agree that it makes sense to combine the constituent S2G linking strategies, and we agree that how to best combine them is not trivial. The reviewer has raised two specific concerns: (i) use of a linear combination of S2G strategies with genome-wide weights may not optimize performance at a given locus; and (ii) our cS2G strategy does not make use of knowledge of relevant tissues/cell types/cell states for some diseases/traits. Below, we discuss each of these concerns in turn.

We note that previously proposed combined strategies (GeneHancer and Open Targets) suffer both of these limitations, and underperform cS2G (precision = 0.75 for cS2G, 0.14 for GeneHancer, 0.33 for Open Targets; Supplementary Table 5). We also note our new findings on the promising performance of cS2G on experimentally validated SNP-gene pairs (see response to Reviewer #2 Comment 2).

Open Access This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

(i) use of a linear combination of S2G strategies with genome-wide weights may not optimize performance at a given locus.

We have updated the Discussion section (p. 20) to note this limitation of our method, and to state that exploring ways to optimally use locus-specific information is a promising direction of future research.

(ii) our cS2G strategy does not make use of knowledge of relevant tissues/cell types/cell states for some diseases/traits.

The reviewer is correct that our cS2G strategy does not make use of this information, as we included all available tissues and cell types for the constituent S2G strategies of cS2G. This choice was motivated by the result that in analyses of 11 autoimmune diseases and blood cell traits—for which it is expected that blood and immune cell types would be most relevant—including all available tissues and cell types attained higher precision than restricting to blood and immune cell types (Supplementary Figure 5, formerly Supplementary Figure 3), perhaps due to limited biosample size. We agree that exploring ways to optimally use tissue/cell type/cell state-specific S2G links is a promising direction of future research. We have updated the Discussion section (p. 19) to clarify these points.

2. I find it difficult to interpret the results on the UK Biobank. How can we know those 7,111 SNP-gene-disease triplets are “high-confidence” results? Fine-mapping at any single GWAS locus is generally challenging and there are many caveats to such results, so how can we confidently identify so many disease genes in one analysis?

The reviewer has raised 2 related questions: (i) how can we assign high confidence to the 9,670 fine-mapped SNP-disease pairs, given the challenges and caveats of fine-mapping?; and (ii) how can we assign high confidence to the 7,111 SNP-gene-disease triplets? We address each question in turn.

(i) how can we assign high confidence to the 9,670 fine-mapped SNP-disease pairs, given the challenges and caveats of fine-mapping?

The peer-reviewed and published paper of Weissbrod et al. 2020 Nat Genet (ref. 7) demonstrated that functionally informed fine-mapping under the PolyFun framework robustly identifies a very large number of fine-mapped SNPs for UK Biobank diseases/traits while producing well-calibrated posterior inclusion probabilities (PIP). This enables us to assign high confidence to the 9,670 fine-mapped SNP-disease pairs (PIP > 0.50). We have updated the *Leveraging the combined S2G strategy to pinpoint disease genes* subsection of the Results section (p. 14) to clarify this point. We believe that revisiting the analyses of Weissbrod et al. demonstrating the robustness of its methods is outside the scope of this manuscript.

Open Access This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

(ii) how can we assign high confidence to the 7,111 SNP-gene-disease triplets?

First, we agree that quantifying the confidence of each individual SNP-gene-disease triplet is very important. We have modified the manuscript to place greater emphasis on the confidence score of each triplet (see response to Reviewer #2 Comment 1), specifically emphasizing the 5,095 SNP-gene-disease triplets with confidence score > 0.50. We have modified the Abstract (p. 2), *Leveraging the combined S2G strategy to pinpoint disease genes* subsection of the Results section (p. 15), and Discussion section (p. 18) accordingly.

Second, we have extensively validated our cS2G strategy (also see response to Reviewer #2 Comment 2):

We have added new analyses of 17 disease/trait loci (including 12 loci reported in the review paper of Gallagher et al. 2018 Am J Hum Genet; ref. 58) containing 25 experimentally validated causal SNP-gene pairs that were previously validated using a combination of functional data and experimental follow-up. Of the 25 experimentally validated causal SNP-gene pairs, 16 had the causal SNP annotated with cS2G linking score >0.5, of which 11 had the causal SNP accurately linked to the validated causal gene. We thus estimated precision as 11/16 = 0.69 (s.e. = 0.12), and recall as 11/25 = 0.44 (s.e. = 0.10). This is a lower precision and higher recall than our estimates based on validation critical gene sets (0.75 for precision, 0.33 for recall), but the differences were not statistically significant due to the small number of experimentally validated SNP-gene pairs. We note that cS2G obtained the 2nd best precision and the best recall when compared to its constituent strategies (max precision = 0.78, s.e. = 0.14 for EpiMap), and higher precision than the Closest TSS strategy (0.56, s.e. = 0.10), but differences were not significant due to the noise of the N=25 sample size. We note that the *FTO*/*IRX3*/*IRX5* locus mentioned by the reviewer was included in these analyses, but the causal SNP rs1421085 was not annotated with cS2G linking score >0.5, as it was linked to 2 genes (*FTO* and *RPGRIP1L*) each with a score of 0.5, and we decided (prior to these analyses; see originally submitted manuscript) to prioritize SNPs linked to a gene with cS2G linking score >0.5 (if we had instead used cS2G linking score ≥0.5 as our threshold, the impact on results would be minimal, with precision = 11/17 = 0.65 instead of 0.69). Overall, we consider this to be a promising validation of the potential of cS2G to identify causal SNP-gene pairs. We have included these new results in a new *Validating the combined S2G strategy using curated lists of disease-associated SNP-gene pairs* subsection of the Results section (p. 12), citing a new Table 2 and new Supplementary Tables 12 and 13.

To further support our cS2G predictions, we previously analyzed a curated disease-associated list of 577 linked sentinel SNP-gene pairs with the underlying genes validated with high confidence by Open Targets (now published in Mountjoy et al. 2021 Nat Genet; ref. 42), with the caveat that these do not represent experimentally validated SNP-gene pairs. We obtained a precision of 0.58 and recall of 0.36. We note that these results are impacted by the fact that Open Targets reports sentinel SNPs rather than causal fine-mapped SNPs (see Supplementary

Open Access This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

Table 16, formerly Supplementary Table 17, for verification on UK Biobank fine-mapping analyses). These results are now reported in the new *Validating the combined S2G strategy using curated lists of disease-associated SNP-gene pairs* subsection of the Results section (p. 12).

We have now analyzed a separate curated disease-associated list of 1,668 sentinel SNP-gene pairs validated using nearby fine-mapped protein-coding SNPs (Weeks et al. medrxiv), again with the caveat that these do not represent experimentally validated SNP-gene pairs. We observed a precision of 0.66 and recall of 0.46. We note that these results are impacted by the fact that the Weeks et al. curated list assumes that a gene with a protein-coding fine-mapped SNP is always targeted by 1Mb surrounding non-coding fine-mapped SNPs. We have updated the new *Validating the combined S2G strategy using curated lists of disease-associated SNP-gene pairs* subsection of the Results section to note these results (p. 12).

3. Similar comments about the results on checking the omnigenic model. Those are quite speculative in nature, and difficult to verify.

We are unsure which specific limitation of the assessment of disease omnigenicity the reviewer is referring to. We do feel that the very distinct patterns of disease omnigenicity inferred using the cS2G vs. Closest TSS strategies (Figure 5a) strongly support the use of functionally informed S2G strategies in such analyses. However, we agree that our results are subject to limitations (e.g. our analyses do not capture *trans* effects) and difficult to verify. We have updated the Discussion section (p. 20) to clarify these points.

Minor comment:
4. The paper is very densely written, which interferes with the flow of the paper.

We agree that the paper is densely written, as it includes a lot of content. We have added a new Results subsection titled "*Validating the combined S2G strategy using curated lists of disease-associated SNP-gene pairs"*, containing both old and new analyses, in an effort to limit the number of analyses included in each individual Results subsection. We have also moved the large amount of text interpreting Figure 4a-d into a new Supplementary Note, while retaining Figure 4 as a main Figure. We are very open to further expanding the text to make the paper less dense, if the editor communicates that a less dense paper with a larger word count would be preferred. We are also open to moving some of the analyses to the Supplementary material, although we feel that the new analyses suggested by the reviewers are important.

Decision Letter, second revision:

Open Access This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

nature portfolio

Our ref: NG-AN57945R2

26th Jan 2022

Dear Steven,

Thank you for submitting your revised manuscript "Combining SNP-to-gene linking strategies to pinpoint disease genes and assess disease omnigenicity" (NG-AN57945R2). 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,

Michael Fletcher, PhD
Associate Editor, Nature Genetics

ORCID: 0000-0003-1589-7087

Reviewer #1 (Remarks to the Author):

All my comments have be addressed.

My one remaining concern is the clarity of the method description. It was almost as difficult to understand the approach as it was on initial reading. Figure 1 is a good overview, but only mentions precision, whereas the recall and heritability aspects are equally important. Could this figure be extended (or an additional figure added) to help illustrate these concepts?

Reviewer #2 (Remarks to the Author):

I thank the authors for considering and addressing appropriately my comments. I think that emphasizing (more) the relevance of the results as they relate to functional characterization will make

 **Open Access** This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

the manuscript appeal to a wider readership. In particular, the results summarized in Figure 3B and STable 18 are of particular interest in that regard. I also enjoyed the balanced discussion of the few examples of the known SNP-gene-disease triplets, and agree that even if the number is small, results are encouraging.

At this point, I don't have additional comments.

Signed: Guillaume Lettre

Reviewer #3 (Remarks to the Author):

The authors have addressed my previous concerns.

Author Rebuttal, second revision:

Response to reviewers for NG-AN57945R1 (Gazal et al.)

Reviewer #1:

All my comments have been addressed.

My one remaining concern is the clarity of the method description. It was almost as difficult to understand the approach as it was on initial reading. Figure 1 is a good overview, but only mentions precision, whereas the recall and heritability aspects are equally important. Could this figure be extended (or an additional figure added) to help illustrate these concepts?

We thank the reviewer for the appreciation of our efforts.

We agree with the reviewer suggestion to improve the clarity of the method description. Per the reviewer suggestion, we have added a new panel to Figure 1 (Figure 1b) illustrating h^2 coverage, and a sentence in the Figure caption defining recall as the product of the h^2 coverage and precision. We have further improved the clarity of the method description by modifying and shortening the *Overview of methods* subsection of the Results section.

Reviewer #2:

I thank the authors for considering and addressing appropriately my comments. I think that

Open Access This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

emphasizing (more) the relevance of the results as they relate to functional characterization will make the manuscript appeal to a wider readership. In particular, the results summarized in Figure 3B and STable 18 are of particular interest in that regard. I also enjoyed the balanced discussion of the few examples of the known SNP-gene-disease triplets, and agree that even if the number is small, results are encouraging.

At this point, I don't have additional comments.

Signed: Guillaume Lettre

We thank the reviewer for validating our efforts and for the accurate summary of our additional analyses and edits to the text, and for confirming that they don't have additional comments.

Reviewer #3:

The authors have addressed my previous concerns.

We thank the reviewer for indicating that their concerns have been addressed.

Final Decision Letter:

In reply please quote: NG-A57945R3 Gazal

27th Apr 2022

Dear Steven,

I am delighted to say that your manuscript "Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity" 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.

Open Access This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

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-A57945R3) 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 Nature Genetics 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. Find out more about Transformative Journals

Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) 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, the journal's standard licensing terms will need to be accepted, including <a href="https://www.nature.com/nature-portfolio/editorial-policies/self-archiving-and-license-to-publish. Those 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 Portfolio offers an immediate open access option only for papers that were

Open Access This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

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 https://www.nature.com/reprints/author-reprints.html. 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-A57945R3). Further information can be found at https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards#protocols

Sincerely,

Michael Fletcher, PhD
Senior Editor, Nature Genetics

ORCiD: 0000-0003-1589-7087

Open Access This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.