nature portfolio

Peer Review Information

Journal: Nature Genetics
Manuscript Title: Polygenic enrichment distinguishes disease associations of individual cells in single-cell RNA-seq data
Corresponding author name(s): Dr. Martin Zhang

Reviewer Comments & Decisions:

Decision Letter, initial version:

11th Nov 2021

Dear Martin,

Your Technical Report, "Polygenic enrichment distinguishes disease associations of individual cells in single-cell RNA-seq data" has now been seen by 2 referees. You will see from their comments below that while they find your work of interest, some important points are raised. We are interested in the possibility of publishing your study in Nature Genetics, but would like to consider your response to these concerns in the form of a revised manuscript before we make a final decision on publication.

Briefly, the two reviewers appreciate the potential of scDRS. Reviewer #1 is very positive; their comments are focused on areas that would better communicate the power of the method, and its potential limitations. Reviewer #2, while acknowledging that scDRS has improved power compared to extant methods, thinks that the biological novelty presented does not, currently, make a convincing case for its use.

We note that both reports draw attention to an important aspect of scDRS: the definition of disease-associated genes from GWAS results. Reviewer #1 asks how GWAS of variable power will affect the results of scDRS; Reviewer #2 asks why a binary definition of a disease-associated geneset is used, instead of a more sophisticated approach. We think it would be important to fully clarify and justify the choice of the top 1000 associated genes, and that the suggestions of Reviewer #2 for different definitions could be interesting to explore further in a revision, as they suggest it may lead to further biological discoveries.

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 Technical Report format instructions, available
<a href="http://www.nature.com/ng/authors/article_types/index.html">here</a>.
Refer also to any guidelines provided in this letter.

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

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

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

[REDACTED]

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

We hope to receive your revised manuscript within six months. If you cannot send it within this time, please let us know. Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

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

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

Sincerely,

Michael Fletcher, PhD
Associate Editor, Nature Genetics

ORCID: 0000-0003-1589-7087

Referee expertise: both reviewers are experts in statistical genetics.

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
This paper introduces an important and interesting question - can we use GWAS associations to determine the specific cell types where causal action underlying disease risk is enacted. The new method, scDRS, generates a test statistic for each individual cell for a given GWAS trait. The method has many steps, including some heuristic steps. The steps are logically presented, seem reasonable and are tested first with simulation, and key results are replicated with different data sets where possible. The method identifies the top 1000 genes for a GWAS trait and then matches the gene expression of these genes with those expressed in individual cells, identifying cells and cell types where the GWAS-genes are more frequently found compared to random sets of genes with similar mean/variance properties. The method is novel and an important advance on existing methods.

There is a lot in the paper! Some of the analyses conducted with the cell sub-populations are like papers in their own right. My initial set of comments was much longer, but as I read and re-read I could answer the comments myself. On final read, I found the paper to be very clear. My remaining comments will hopefully help improve readability.

Main comments
1) On my first read it took a while for me to really appreciate that scDRS uses individual cells rather than individual cell types. When I re-read with this knowledge it was hard to see why I had missed this (!), and the sentences that seemed untrue at first reading such as "To our knowledge, scDRS is the first method to associate individual cells in scRNA-seq data to disease GWAS." made sense. I wonder if you can signpost this better through added emphasis, since after all the first application X has dimension number of pre-defined cell types

Eg abstract insert “without the need for annotation of individual cells to cell-types” after “with polygenic risk of disease at individual cell resolution”
Line 54 “e.g., individual cells” update “e.g., individual cells both within and across cell-types”

2) in trying to understand how scDRS compares with other methods, I noted that ref 7 contrasts their approach to ref 26, then ref 8 which is from the same authors as ref 26 provide detailed reasons (First page of methods second column) why the assumptions made by ref 7 are not valid. Some of the concerns raised about assumptions imposed by ref 7 are also assumptions made here. I would have liked to come away with a clearer view of the opinion of the authors about the benefits of scDRS over the method in ref 8 for the cell-type analysis. Perhaps this could be added to the supplement. Specifically, I think ref 8 argues that their method does not focus on genes that are highly expressed per se in any cell type but on genes that show cell-type specificity in expression even though absolute level of expression may not be high. I believe in scDRS the use of control sets speaks to this point, but I would feel more confident if the authors gave their opinion

3) I am unclear about the impact of power in the GWAS in the results. The traits studied have “heritability z-score>6”, but still there are differences in power between data sets. Not all sets of 1000 disease genes are equally powered (in addition to the comparison in supp T15, where optimal number of genes to select is likely a reflection of polygenicity).Take for example, Fig 4B (IBD =CD+UC) shows more red than Supp Fig 14 A (CD) and B (UC). Maybe the differences between traits in Supp Fig14 reflects power. The authors should provide guidelines in the discussion about when GWAS data are not sufficiently powered for this analysis.

Minor comments.

Line 89 “scDRS outputs individual cell-level p-values, normalized disease scores, and 1,000 sets of normalized control scores”
Might be less ambiguous to explain the test for the p-values

Line 240: “we verified that cells from different tissues, age, or sex clustered together”
I wasnt sure if there was a “not” needed, since sex at least is not associated with cluster, and if the not is not required, then it is implied that this is a good thing which should be explained.

Line 331. Seems strange not to include bipolar disorder (and if you do use latest Mullins 2021).

Line 404. If the GWAS associations for metabolic traits are associated with ploidy level does that not imply there should be SNPs (or other DNA variants) associated with ploidy (ploidyQTL) - are these known and can they be directly integrated with the GWAS results?

Line 439 After “scDRS does not rely on annotations of classical cell types based on known marker genes” add for emphasis, “because the scDRS analysis uses the gene expression levels measured in individual cells”.

Line 489 Emphasise ncell is the total number of cells in the data set to avoid reader confusing this with number of cell-types. Although if I understand correctly the first application Ncell is in fact Ncell-type

Lines 488-493. Ngene is the same for all cells in a data. Can you provide some summaries to help give intuitive feel of X. Provide distributions of proportion of zero values, means, variances, technical variances ?

Line 509-511 to avoid ambiguity clearly state what the mean and variance for a gene is calculated across – estimated across the columns of X. Similarly supplementary note line 1093/1094 state what the mean and variance is calculated across to avoid ambiguity.

Line 512 Would be good to give some intuition to this technical variance. I think it is that under the assumption of a consistent relationship between mean and variance of gene expression, genes with high technical variance have higher variance across cells than expected by this relationship? Can you provided the tech variance per gene per data set as a supplementary data file to allow readers to explore factors that might be associated with this. Is there a correlation of technical variance of genes across data sets?

Line 520. Again to avoid ambiguity "Since the control genes match the mean expression and expression variance of the disease genes" add "across cells".

Supp T1. The variable N format seems unusual -check?

Supp Fig 4. Do the results for FEV1/FVC make sense?

Supp Fig 14 reverse legend order to make easier to follow. The color repetition doesn't help.

Reviewer #2:
Remarks to the Author:
The authors introduced a novel approach to evaluate association of diseases/traits with cell type specific gene expression. The authors performed thorough analyses including simulation and alternative way of scoring the disease association, and it is convincing that the scDRS has more power than other methods. However, I did not see obvious advantage of using the proposed method over existing approaches as findings are mostly consistent and I do not think there are many novel biological findings reported. It is also concerning that the method requires defining binary disease genes by arbitrary selecting the top associated genes, regardless of the strength of the association or the genetic architecture of the diseases/traits. I also think the manuscript can be better organized to make easier for readers to follow. Please see more detailed comments below.

1. The authors defined 1000 top associated genes based on MAGMA as putative causal genes. As the authors also concluded that optimal number of genes depends on the polygenicity of the traits, why did the authors chose to use fixed number of genes for all traits rather than using the optimum number of genes? Did you also try with P-value thresholding? I am not convinced that taking top 1000 genes should be default for all traits. In addition, since MAGMA provides association statistics for all

genes tested, why not use the quantitative weight rather than using binary status?
2. As the authors mentioned, associating the gene expression at single cell level with diseases or traits is unique to previous studies and it is advantageous to be able to find heterogeneity within the known cell type clusters. However, at the single cell level, I would also expect to see some differences across traits from the same domain. For example, for brain related traits, we already know they are genetically highly correlated but by using single cell level expression, I'd expect to see some differences between traits that might explain specificity of the trait rather than shared biology. However, the results presented in the manuscript does not seem to show that. For example, in Fig 3., the association pattern is very similar between SCZ and smoking. Same is true for Supple Fig 19. This might be because the top 1000 genes may be highly overlapping between traits. Have you checked how many genes (of 1000) are overlapping between traits in the same domain? Isn't it better to use quantitative distribution of gene association so that you might be able to capture subtle difference between traits?
3. Since scDRS uses single cell level expression rather than summarizing by the cell type labels, the distribution of cell types in the dataset can bias the mean expression of the genes. For example, for a brain scRNA-seq dataset, neuronal cells might be dominant and the average expression of a gene is basically the average expression of neuronal cells rather than cells in the brain. The authors should discuss potential issues regarding the unbalanced dataset.
4. There is no mention about the computational cost and expected run time of this method. Is it scalable with the number of cells in the dataset? I see the largest dataset used in this study has 500K cells, but can you run this method for > 1 million cells in a reasonable time? For example, Tabula Sapiens has 2 million cells in the dataset.
5. I do think the authors have done great amount of work and performed comprehensive analyses but I found that the manuscript is a little hard to follow. I understand that there are many results to show but, especially for such high impact journal with broad readers, the authors should spend little more effort to make the main objectives and findings clearer. In addition, there are many repetitive expressions, which does not make it easier to read. It might be better to focus on main findings in the main text and move some of the analyses into supplementary information. For example, line 210-231 can be summarized in a few sentences and move details to supplementary information. In the 3 sections describing results of heterogeneous subpopulations of associated cells for autoimmune, brain-related and metabolic traits, the secondary analyses can be replaced with a few sentences and move details to supplementary information.

Author Rebuttal to Initial comments

Response to reviewers for NG-TR58415R

Reviewer #1

Remarks to the Author:
This paper introduces an important and interesting question - can we use GWAS associations to determine the specific cell types where causal action underlying disease risk is enacted. The new method, scDRS, generates a test statistic for each individual cell for a given GWAS trait. The method has many steps, including some heuristic steps. The steps are logically presented, seem reasonable and are tested first with simulation, and key results are replicated with different data sets where possible. The method identifies the top 1000 genes for a GWAS trait and then matches the gene expression of these genes with those expressed in individual cells, identifying cells and cell types where the GWAS-genes are more frequently found compared to random sets of genes with similar mean/variance properties. The method is novel and an important advance on existing methods.

There is a lot in the paper! Some of the analyses conducted with the cell sub-populations are like papers in their own right. My initial set of comments was much longer, but as I read and re-read I could answer the comments myself. On final read, I found the paper to be very clear. My remaining comments will hopefully help improve readability.

We thank the reviewer for the accurate summary of our method and for noting the novelty of our approach as an important advance on existing methods. Indeed, we agree with the reviewer in that our manuscript is extensive and includes a lot of results; in revision we attempted to reduce the text to improve readability and highlight the main points better.

Main comments
1) On my first read it took a while for me to really appreciate that scDRS uses individual cells rather than individual cell types. When I re-read with this knowledge it was hard to see why I had missed this (!), and the sentences that seemed untrue at first reading such as “To our knowledge, scDRS is the first method to associate individual cells in scRNA-seq data to disease GWAS.” made sense. I wonder if you can signpost this better through added emphasis, since after all the first application X has dimension number of pre-defined cell types

Eg abstract insert “without the need for annotation of individual cells to cell-types” after “with polygenic risk of disease at individual cell resolution”
Line 54 “e.g., individual cells” update “e.g., individual cells both within and across cell-types”

We agree that emphasizing the fact that scDRS identifies disease associations of individual cells rather than individual cell types is of high importance. We have updated the Abstract (p.2) and Introduction (p.2) using the text suggested by the reviewer as *“…with polygenic risk of disease at individual cell resolution <u>without the need for annotation of individual cells to cell types</u>”* and *“individual cells <u>both within and across cell types</u>”*, respectively.

We have also updated other parts of the Introduction section (p.2) as *“…overlooks the considerable heterogeneity <u>of individual cells</u> within cell types…”* and updated the *Overview of methods* subsection of the Results section (p.2) as *“…associate individual cells to disease <u>without the need for annotation of individual cells to cell types</u>…”.*

2) in trying to understand how scDRS compares with other methods, I noted that ref 7 contrasts their approach to ref 26, then ref 8 which is from the same authors as ref 26 provide detailed reasons (First page of methods second column) why the assumptions made by ref 7 are not valid. Some of the concerns raised about assumptions imposed by ref 7 are also assumptions made here. I would have liked to come away with a clearer view of the opinion of the authors about the benefits of scDRS over the method in ref 8 for the cell-type analysis. Perhaps this could be added to the supplement. Specifically, I think ref 8 argues that their method does not focus on genes that are highly expressed per se in any cell type but on genes that show cell-type specificity in expression even though absolute level of expression may not be high. I believe in scDRS the use of control sets speaks to this point, but I would feel more confident if the authors gave their opinion

The reviewer has raised 2 questions, citing ref. 7 (Watanabe et al. 2019 Nat Commun), ref. 8 (Bryois et al. 2020 Nat Genet), and ref. 26 (Skene et al. 2018 Nat Genet): (a) do the concerns of ref. 8 about the assumptions of ref. 7 also apply to scDRS; and (b) what are the benefits of scDRS over the method of ref. 8 for the cell type-level analysis. We address each question in turn. We focus the MAGMA-based methods in both ref. 8 and ref. 26, because they are more comparable to scDRS.

(a) Do the concerns of ref. 8 about the assumptions of ref. 7 also apply to scDRS?

We first focus on the concern regarding highly-expressed vs. specifically-expressed genes, and clarify that scDRS also focuses on specifically-expressed genes like ref. 8. Specifically, ref. 8 stated that ref. 7 *"hypothesize that genes with higher levels of expression should be more associated with a trait"* and raised the concern that *"many cell-type-specific genes that are disease relevant are expressed at moderate levels"*. scDRS individual cell-level analysis is not susceptible to this concern because it assesses the excess expression of a disease-associated gene in each cell relative to that of a control gene that matches the mean expression of the disease-associated gene across cells. Since scDRS cell type-level analysis is a direct result of the individual cell-level analysis, it is not susceptible to this concern either. In this respect, scDRS is more similar to ref. 8, which assesses the expression of a gene in a cell type relative to the total expression across cell types (defined as specificity in ref. 8). We have updated the *Overview of methods* subsection of the Results section (p.2), the Methods section (p.13), and the Supplementary Note (p.s5) to clarify this point.

For completeness, we next discuss scDRS in the context of other concerns that ref. 8 raised about ref.7 and that ref. 7 raised about ref. 26.

- Ref. 8 noted that the method in ref. 7 depends on the set of cell types in the data set. We believe that this concern applies to most existing methods, including ref. 7, including ref. 8 and ref. 26 (because in ref. 8 and ref, 26, the cell type specificity of a given gene is defined relative to the total expression of that gene in all cell types within the analyzed data set), and including scDRS. We believe that this concern is best addressed by making clear that results should be interpreted with respect to other cells (or cell types) in the data set. We have expanded our discussion of this point in the Discussion section (p.9) and updated the Methods section (p.13), and the Supplementary Note (p.s5).
- Ref. 8 also noted that the method in ref. 7 is sensitive to scaling (scaling number of counts per cell in scRNA-seq to the same number). This concern applies to scDRS and ref. 7 but do

not apply to ref. 8 and ref. 26. We note that preprocessing single-cell data with size factor normalization and log transformation (as performed in scDRS) is a recommended best practice for single-cell data preprocessing and is essential for correcting for the confounding effect of sequencing depth and stabilizing the count data (Luecken Mol Syst Biol 2019, ref. s12; Stuart Cell 2019, ref. 21). Nonetheless, we determined that scDRS produced consistent individual cell-level results with different scaling factors (median score correlation 0.90 across 74 traits between scaling to the default 10,000 vs. 1 million reads per cell). The cell type-level results are a direct consequence of the individual cell-level results and are expected to be also consistent. We have updated the *Results across 120 TMS cell types …* subsection of the Results section (p.5), the Methods section (p.13), and the Supplementary Note (p.s4, p.s5) to discuss this potential concern and note this reassuring result.

- Ref. 7 raised the concern that ref. 26 did not condition on the average expression level of a gene across cell types when regressing the MAGMA z-scores against cell type specificity across genes, and is thus "vulnerable to confounding by a general effect of gene expression". This concern also applies to ref. 8. scDRS is not susceptible to this concern, because a general effect of gene expression on trait association would be present in both the disease-associated genes and the matched control genes (which have the same mean expression), and would thus cancel each other out. We have updated the Supplementary Note (p.s5) to clarify this point.

More generally, ref. 7,8,26 use linear regression to associate MAGMA z-scores with cell type features across genes (cell type expression level in ref. 7; cell type specificity in ref. 8,26). scDRS can be viewed as a non-parametric alternative to these methods, employing a stratified permutation test that associates MAGMA z-scores for top genes (we additionally added MAGMA z-score gene weights to improve power; see part (b) of response to Reviewer #2 Comment 1) with expression levels for a given cell by permuting genes within levels of expression mean and variance. Thus, unlike ref. 7,8,26, scDRS does not rely on a linearity assumption. We have updated the Methods section (p.13) and the Supplementary Note (p.s5) to discuss these points.

(b) What are the benefits of scDRS over the method of ref. 8 for the cell type-level analysis?

Although the main innovation of scDRS pertains to disease associations of individual cells, we agree that further dissection and discussion of cell type-level analyses is warranted.

Motivated by reviewer comment we expanded the comparison of cell type-level analyses to include the 3 approaches used in ref. 8: LDSC using the top 10% high-specificity genes (instead of differentially expressed genes, as in LDSC-SEG), MAGMA using the top 10% high-specificity genes, and the combined method (mean -log10 p-values of the first two methods). Similar to previous analyses comparing to LDSC-SEG, we again determined that the results are highly consistent (r=0.69, r=0.64, r=0.7 for association -log10 p-values, respectively). We have updated the *Results across 120 TMS cell types…* subsection of the Results section (p.5), citing Supplementary Figure 11 and the Supplementary Note (p.s4), to include these results.

In addition to comparison with existing methods, we also highlight areas where scDRS has 2 potential advantages over other methods such as MAGMA using the top 10% high-specificity genes as employed in ref. 8. First, scDRS does not assume a linear relationship between MAGMA z-scores and (binarized) cell type specificity across genes (see part (a) of the response). Second, scDRS may be more robust to within-cell type heterogeneity. For example, if only 20% of cells in a heterogeneous cell type are disease-associated, the cell type specificity as defined in ref. 8 will be dominated by the non-associated 80% of cells and may not be able to capture the associated subpopulation. In comparison, scDRS tests if the top 5% most associated cells in the cell type are significantly associated with the disease, and may therefore be able to capture the associated subpopulation. We have updated the Methods section (p.13) and the Supplementary Note (p.s5) to discuss these points, while noting that the primary focus of our work is on disease associations of individual cells and that further investigation of cell type-level analyses remains as a future direction.

3) I am unclear about the impact of power in the GWAS in the results. The traits studied have “heritability z-score>6”, but still there are differences in power between data sets. Not all sets of 1000 disease genes are equally powered (in addition to the comparison in supp T15, where optimal number of genes to select is likely a reflection of polygenicity).Take for example, Fig 4B (IBD =CD+UC) shows more red than Supp Fig 14 A (CD) and B (UC). Maybe the differences between traits in Supp Fig14 reflects power. The authors should provide guidelines in the discussion about when GWAS data are not sufficiently powered for this analysis.

We fully agree that it is of interest to carefully assess the impact of GWAS power on the scDRS results.

To address this, we performed a new experiment to assess the impact of GWAS power on scDRS as follows. We subsampled the UK Biobank GWAS diseases/traits at various sample sizes (10K, 30K, 100K, 300K, in addition to the full set of 459K samples) and reassessed scDRS performance at each of these data sets. We focused on individual cell-level discoveries.

As expected, we found that scDRS discoveries increase with GWAS power and that a GWAS sample size greater than 100K (or heritability z-score greater than 5) is desirable for scDRS to produce a reasonable number of discoveries across diseases/traits (although less stringent thresholds can be used for less polygenic traits).

We changed the heritability z-score threshold from 6 to 5 to be consistent with this experiment, although this change does not have a major impact on the results. We have updated the Discussion section (p.8, citing new Supplementary Figure 35) to provide these results and state that we recommend that scDRS be run on GWAS data sets with a heritability z-score greater than 5, or sample size greater than 100K if heritability z-score is not available (although less stringent thresholds can be used for less polygenic traits). We further note that, for less well-powered GWAS data sets, scDRS will not produce false positives, but may produce less significant results.

Minor comments.

Line 89 “scDRS outputs individual cell-level p-values, normalized disease scores, and 1,000 sets of normalized control scores”
Might be less ambiguous to explain the test for the p-values

We have revised the text in the *Overview of methods* subsection of the Results section (p.3) as follows: *“scDRS outputs individual cell-level p-values (testing for cell-disease association as described above), normalized disease scores, and 1,000 sets of normalized control scores”.*

Line 240: “we verified that cells from different tissues, age, or sex clustered together”

I wasnt sure if there was a “not” needed, since sex at least is not associated with cluster, and if the not is not required, then it is implied that this is a good thing which should be explained.

Indeed, that was a typo. We meant to show that there were no batch effects due to tissues, age or sex. We have revised the text in the *Heterogeneous subpopulations of T cells…* subsection of the Results section (p.5) as *“we verified that batch effects were not observed for tissue, age, or sex”.*

Line 331. Seems strange not to include bipolar disorder (and if you do use latest Mullins 2021).

We have updated the bipolar summary statistics from Stahl 2019 Nat Genet to Mullins 2021 Nat Genet as suggested by the reviewer, and included bipolar in the main analysis in the *Results across 120 TMS cell types …* subsection (p.4, citing an updated Figure 3) and the *Heterogeneous subpopulations of neurons …* subsection of the Results section (p.6, citing an updated Figure 5B). Of note, scDRS produced similar results for these two sets of summary statistics, identifying diverse brain cell types as being relevant. However, the new bipolar summary statistics from Mullins et al. produced more powerful scDRS results than the old bipolar summary statistics from Stahl et al. (655 vs. 104 for the number of significantly associated cells).

Line 404. If the GWAS associations for metabolic traits are associated with ploidy level does that not imply there should be SNPs (or other DNA variants) associated with ploidy (ploidyQTL) - are these known and can they be directly integrated with the GWAS results?

It is indeed plausible that the association between hepatocyte ploidy level and metabolic traits may imply that there are metabolic trait GWAS variants associated with ploidy (ploidyQTL). However, this is difficult to assess directly, as genetic studies of ploidy level have largely focused on organisms other than humans (e.g., Meng et al. 2016 BMC Plant Biol, ref. 84), perhaps because most human cells are diploid. We have updated the *Heterogeneous subpopulations of hepatocytes…* subsection of the Results section (p.8), citing the Supplementary Note (p.s6), to note this point.

Instead, we considered a published polyploidy signature gene set (Miettinen et al. 2014 Curr, ref. 81; used for computing the polyploidy score in our paper), and investigated whether the GWAS gene sets for metabolic traits (from MAGMA) were enriched for polyploidy signature genes. We indeed observed significant enrichment (average odds ratio of 1.46 with SE=0.09 across the 9 metabolic traits), suggesting that GWAS SNPs for the 9 metabolic traits may regulate expression levels of these polyploidy genes and therefore may be enriched for ploidyQTL. We have updated the *Heterogeneous subpopulations of hepatocytes…* subsection of the Results section (p.8), citing the Supplementary Note (p.s6), to note this result.

Finally, we have updated the Discussion section (p.9) to note the potential of incorporating eQTL data, which are much more broadly available then ploidyQTL data.

Line 439 After “scDRS does not rely on annotations of classical cell types based on known marker genes” add for emphasis, “because the scDRS analysis uses the gene expression levels measured in individual cells”.

We have revised the text in the Discussion section (p.8) as “*scDRS does not rely on annotations of classical cell types based on known marker genes, a standard approach for integrating GWAS with scRNA-seq data (and bulk gene expression data; see Supp. Note), because the scDRS analysis uses the gene expression levels measured in individual cells.*”

Line 489 Emphasise ncell is the total number of cells in the data set to avoid reader confusing this with number of cell-types. Although if I understand correctly the first application Ncell is in fact Ncell-type

We have revised the text in the Methods section (p.9) as “*We consider a scRNA-seq data set with n_cell cells (not cell types) and n_gene genes.*” We further note the analyses in the first application (the *Results across 120 TMS cell types…* subsection of the Results section) was first performed at individual cell-level and then aggregated at cell type-level, so n_cell still refers to the number of cells (not cell types). We have revised the text as “*We first report scDRS cell type-level results, aggregated for each cell type from the scDRS individual cell-level results…*” to clarify this point (p.4).

Lines 488-493. Ngene is the same for all cells in a data. Can you provide some summaries to help give intuitive feel of X. Provide distributions of proportion of zero values, means, variances, technical variances ?

We computed the distribution of zero proportion, mean expression, expression variance, and technical variance across genes for the TMS FACS, TMS Droplet, and TS FACS data. These distributions are similar to exponential distributions from visual inspection. Results are reported in the Methods section (p.9), citing new Supplementary Figure 2 and new Supplementary Table 3.

Line 509-511 to avoid ambiguity clearly state what the mean and variance for a gene is calculated across – estimated across the columns of X. Similarly supplementary note line 1093/1094 state what the mean and variance is calculated across to avoid ambiguity.

We have revised the text in the Methods section (p.9) as *“...by randomly selecting genes matching the mean expression and expression variance of the disease genes calculated across all cells in the data set…”* and text in the Supplementary Note (p.s2) as “*Specifically, we first compute the mean expression and expression variance for each gene across all cells in the data set in the original non-log-transformed space.*”

Line 512 Would be good to give some intuition to this technical variance. I think it is that under the assumption of a consistent relationship between mean and variance of gene expression, genes with high technical variance have higher variance across cells than expected by this relationship? Can you provided the tech variance per gene per data set as a supplementary data file to allow readers to explore factors that might be associated with this. Is there a correlation of technical variance of genes across data sets?

The technical variance of a gene in a single-cell data set is defined as the part of gene expression variance across cells that is due to the technical sequencing noise, rather than the true biological variation across cells. We used a commonly-used estimator for technical variance (Seurat log-normalization; Frost 2021 Nucleic Acids Res, ref. 18; Stuart 2019 Cell, ref. 21), which estimates the technical variance as the proportion of the variance that can be predicted by the mean expression via modeling the mean-variance relationship across genes. We have revised the text in the Methods section (p.9) to clarify this as *“...Next, we estimate the technical noise level for each gene in scRNA-seq data, the part of the variance due to sequencing noise, using a procedure similar to previous works by modeling the mean-variance relationship across genes; we further compute the raw disease score…”.*

We now report the technical variance and other statistics (mean, variance, zero proportion) for each gene and each of the 16 data sets in Supplementary Table 3, cited in the Methods section (p.9). We also report the correlation of the technical variance across genes between the 16 data sets in Supplementary Table 4, cited in the Methods section (p.9). The correlations are generally positive across data sets (mean 0.34 and SD 0.23), and are substantially higher for data sets with similar cell type compositions (e.g., 0.8 between TMS FACS and TMS droplet, and 0.74 between TMS FACS and TS FACS).

Line 520. Again to avoid ambiguity "Since the control genes match the mean expression and expression variance of the disease genes" add "across cells".

We have revised the text in the Methods section (p.10) as *"Since the control genes match the mean expression and expression variance of the disease genes across cells, …"*

Supp T1. The variable N format seems unusual -check?

We have updated the variable N format in Supplementary Table 1.

Supp Fig 4. Do the results for FEV1/FVC make sense?

We have investigated this further, and believe that the results for FEV1/FVC make sense overall. Specifically, we identified 20 cell types associated with FEV1/FVC (FDR<0.05), including 5 lung cell types and 15 cell types from other tissues:

1. Type II pneumocyte (lung).
2. Skin-related: basal cell (mammary gland), bulge keratinocyte (skin).
3. Smooth muscle cells: smooth muscle cell (heart), smooth muscle cell of trachea (trachea), valve cell (heart).
4. Fibroblast-and-MSC-like (mesenchymal stem cell): fibroblast of lung (lung), pulmonary interstitial fibroblast (lung), fibroblast (trachea), fibroblast of cardiac tissue (heart), pancreatic stellate cell (pancreas), mesenchymal stem cell (limb muscle / diaphragm), mesenchymal stem cell of adipose (fat tissues including SCAT, BAT, GAT, MAT), stromal cell (mammary gland), chondrocyte (trachea), bladder cell (bladder).
5. Pericyte-like: pericyte cell (lung), adventitial cell (lung), brain pericyte (non-myeloid brain), mesangial cell (kidney).

The first 4 sets of associations are consistent with previous work which reported type II pneumocytes, basal cells, smooth muscle cells, and fibroblasts to be associated with chronic obstructive pulmonary disease (COPD), a lung disease closely related to FEV1/FVC (Sakornsakolpat et al. 2019 Nat Genet, ref. s70). The pericyte association is also plausible because pericytes are known to regulate lung morphogenesis (Kato et al. 2018 Nat Commun, ref. s71). We note that the cell type associations from the lung are more likely to be causal and those from other tissues are more likely tagging the causal cell types due to shared expression. We have updated the caption of Supplementary Figure 7 to clarify these points.

Supp Fig 14 reverse legend order to make easier to follow. The color repetition doesn't help.

We have reversed the legend order and removed the color repetition (Supplementary Figure 15).

Reviewer #2

Remarks to the Author:
The authors introduced a novel approach to evaluate association of diseases/traits with cell type specific gene expression. The authors performed thorough analyses including simulation and alternative way of scoring the disease association, and it is convincing that the scDRS has more power than other methods. However, I did not see obvious advantage of using the proposed method over existing approaches as findings are mostly consistent and I do not think there are many novel biological findings reported. It is also concerning that the method requires defining binary disease genes by arbitrary selecting the top associated genes, regardless of the strength of the association or the genetic architecture of the diseases/traits. I also think the manuscript can be better organized to make easier for readers to follow. Please see more detailed comments below.

We thank the reviewer for noting that scDRS is a novel approach that attains higher power than other methods. Reviewer concerns regarding novel biological findings, details of how the method selects and weights disease-associated genes, and manuscript organization are addressed below. In particular, we believe that our new results on differences across diseases/traits from the same domain (e.g., new Figure 4C-D; see part (a) of response to Reviewer #2 Comment 2) constitute an important addition to the manuscript and increase biological novelty.

1. The authors defined 1000 top associated genes based on MAGMA as putative causal genes. As the authors also concluded that optimal number of genes depends on the polygenicity of the traits, why did the authors chose to use fixed number of genes for all traits rather than using the optimum number of genes? Did you also try with P-value thresholding? I am not convinced that taking top 1000 genes should be default for all traits. In addition, since MAGMA provides association statistics for all genes tested, why not use the quantitative weight rather than using binary status?

The reviewer has raised 2 related questions: (a) why does scDRS always use the top 1,000 genes, instead of using the optimum number of genes or p-value thresholding; and (b) why not use quantitative gene weights? We address both questions using an extensive comparison across 6*4=24 combinations of gene selection and weights: 6 methods for selecting putative disease genes: top 100, top 500, top 1,000, top 2,000, FDR<1%, and FWER<5% genes (with the number of top genes constrained between 100 and 2,000 for the last 2 methods) and 4 methods for selecting gene weights for the disease genes: unweighted, weights based on MAGMA z-score (capped at 10), single-cell variance-stabilization weights (as before), and weights based on both the MAGMA z-score and variance-stabilization weights. Our use of the MAGMA z-score is a heuristic; it is unclear how to specify theoretically principled gene weights, which must account for complex mechanisms of how genetic variants impact disease through gene expression.

We applied these 24 versions of scDRS to subsampled TMS FACS data sets (20 repetitions with 10K cells each) and a curated set of 20 traits with expected and unexpected disease-critical cell types (Ulirsch 2019 Nat Genet, ref. s34; Finucane et al. 2018 Nat Genet, ref. 12; Ben-Moshe et al. 2019 Nat Rev Gastroenterol Hepatol, ref. 79; Guo et al. 2017 eLife, ref. 38; Chiou et al. 2021 Nat Genet, ref. 41; Verdecchia et al. 2018 Circ Res, ref. s68). We evaluated the 24 versions of scDRS using normalized t-statistics between cells from the expected and unexpected disease-critical cell types; we caution that some cell types labeled as unexpected may still be relevant to disease despite not being implicated in the current literature (Methods).

The version of scDRS that uses the top 1,000 MAGMA genes (as in the original submission) significantly outperformed all other approaches (except top 2,000, which performed similarly to top 1,000). This provides an empirical answer to question (a). In addition, the version of scDRS that uses quantitative gene weights based on the MAGMA z-score (and single-cell variance-stabilization weights, as in the original submission) significantly outperformed all other approaches, confirming the reviewer's intuition that quantitative gene weights are preferred; this provides an empirical answer to question (b). We have updated the *Results across 120 TMS cell types…* subsection of the Results section (p.5, citing Supplementary Figures 12,13, Supplementary Tables 17,18, Supplementary Note (p.s4)) and the Methods section (p.13) to report these findings.

We have updated most analyses in the manuscript to use the new version of scDRS with GWAS MAGMA z-score gene weights (top 1,000 genes; gene weights based on both MAGMA z-score and single-cell variance stabilization weights). Specifically, we have updated Figures 1-5, Supplementary Tables 8,11,12,14-20,22-26, and Supplementary Figures 4,7-14,16,20,24,25,27,28,30,31,34. The new version of scDRS produced highly consistent results with the old version of scDRS (median correlation=0.98 of the scDRS disease scores across the 74 traits). We have updated the Supplementary Note (p.s4), citing Supplementary Figure 13, to note this concordance. However, there are some differences for discoveries close to the detection threshold. For example, 5 out of the 80 significant cell type-disease associations (FDR<0.05) discussed in the current version of Fig. 3 were non-significant in the old version, including GMP and MONO (old and new FDR 0.057 vs. 0.045), Dendritic and RA (0.067 vs. 0.019), Dendritic and MS (0.077 vs. 0.032), Oligodendrocyte and BMI (0.057 vs. 0.045), and Pancreatic beta cell and Smoking (0.1 vs. 0.032). In addition, 5 out of the 70 significant cell type-disease associations discussed in the old version of Fig. 3 were non-significant in the current version, including Pancreatic PP and BMD-HT (old and new FDR 0.046 vs. 0.064), Hepatocyte and RDW (0.046 vs. 0.074), Dendritic and Asthma (0.02 vs. 0.11), Classical monocyte and IBD (0.02 vs. 0.054), and Oligodendrocyte and MDD (0.046 vs. 0.064).

Since the new method involves quantitative GWAS genes weights, we performed additional null simulations to assess the calibration of scDRS with quantitative GWAS genes weights. We used the previous null gene sets and simulated the quantitative GWAS gene weights matching the MAGMA z-score distributions in real traits. We confirmed that scDRS remained well-calibrated. We have updated the *Simulations assessing calibration and power* subsection of the Results section (p.3), Figure 2, and Supplementary Figure 4 to report this result.

Although we have updated most analyses in the manuscript to use the new version of scDRS described above, we note that scDRS can be applied to any set of genes and gene weights supplied by the user. We have updated the Discussion section (p.9) to clarify this point.

2. As the authors mentioned, associating the gene expression at single cell level with diseases or traits is unique to previous studies and it is advantageous to be able to find heterogeneity within the known cell type clusters. However, at the single cell level, I would also expect to see some differences across traits from the same domain. For example, for brain related traits, we already know they are genetically highly correlated but by using single cell level expression, I'd expect to see some differences between traits that might explain specificity of the trait rather than shared biology. However, the results presented in the manuscript does not seem to show that. For example, in Fig 3., the association pattern is very similar between SCZ and smoking. Same is true for Supple Fig 19. This might be because the top 1000 genes may be highly overlapping between traits. Have you checked how many genes (of 1000) are overlapping

between traits in the same domain? Isn't it better to use quantitative distribution of gene association so that you might be able to capture subtle difference between traits?

The reviewer has raised 3 related questions: (a) are there differences across disease/traits from the same domain; (b) what is the overlap of the top 1,000 genes between diseases/traits; (c) do quantitative gene weights help capture differences between diseases/traits. We address each question in turn.

(a) Are there differences across diseases/traits from the same domain?

We computed the scDRS disease score correlation (across the 110,096 TMS FACS cells) for 26 diseases/traits from 3 domains (10 autoimmune diseases, 7 brain traits/diseases, and 9 metabolic traits) considered in the individual cell-level analyses in the paper. The correlations were moderate for diseases/traits from the same domain (avg=0.51 for autoimmune diseases, avg=0.44 for brain traits, avg=0.36 for metabolic traits), implying that there were differences across diseases/traits from the same domain. Furthermore, the 10 autoimmune diseases formed 3 clusters based on hierarchical clustering of scDRS disease score correlations (Supplementary Figure 22): IBD-related (IBD, UC, CD), allergy-related (Eczema, ASM, RR-ENT), and others (MS, RA, AIT, HT). These 3 groups represent biologically more similar diseases within the 10 autoimmune diseases (Yaneva et al. 2021 Asthma Res Prac, ref. 50), suggesting the scDRS results can differentiate between subgroups of diseases from the same domain. Results are reported in the *Heterogeneous subpopulations of T cells …* subsection of the Results section (p.6), citing a new Supplementary Figure 22 and a new Supplementary Table 19.

In addition, we compared the associated cell populations between pairs of representative diseases/traits from the same domain. We first compared IBD (inflammatory bowel disease) with HT (hypothyroidism) over the set of T cells in TMS FACS. We determined that IBD is more associated with cells in cluster 4 (labeled as "Th2/Treg-like") while HT is more associated with other subpopulations, including cells in cluster 3 (labeled as "Treg-like"), cluster 9 (labeled as "CD8+ effector-like"), and cluster 10 (labeled as "proliferative"). These results are reported in new Figure 4C,D and new Supplementary Figure 21. We next compared SCZ (schizophrenia) with Smoking (smoking status) across brain non-myeloid cells in TMS FACS. We determined that oligodendrocytes produced stronger associations for SCZ than for Smoking. These results are reported in new Supplementary Figure 26. Finally, we compared TG (triglycerides) with LDL (low-density lipoprotein) across hepatocytes in TMS FACS. We did not find any notable differences. These results are reported in new Supplementary Figure 33. We have updated the *Heterogeneous subpopulations of T cells…* subsection, the *Heterogeneous subpopulations of neurons…* subsection, and the *Heterogeneous subpopulations of hepatocytes…* subsection of the Results section (p.5, p.7, p.8, respectively), citing Figure 4C,D and Supplementary Figures 21,26,33, to report these results.

We acknowledge that further investigation of differences between diseases/traits within the same domain is an important future direction. We have updated the Discussion section to discuss this point (p.9).

(b) What is the overlap of the top 1,000 genes between diseases/traits?

The gene set overlap is relatively low between diseases/traits. Across the 26 traits from the 3 domains (part (a) of the response), the average overlap of the top 1,000 MAGMA genes is 231 for the 10 autoimmune diseases, 199 for the 7 brain traits, 275 for the 9 metabolic traits, and 111 for pairs of diseases/traits from different domains. Results are reported in the *Heterogeneous* *subpopulations of T cells…* subsection of the Results section (p.6), citing new Supplementary Figure 22 and a new Supplementary Table 19.

Furthermore, the scDRS disease score correlations are not fully driven by the gene set overlaps. We recomputed the scDRS disease score correlations using only the non-overlapping genes of the two gene sets for each pair of diseases/traits. This modified scDRS disease score correlation was still much higher for traits from the same domain (0.16 for autoimmune diseases; 0.17 for brain traits; 0.02 for metabolic traits) than traits from different domains (-0.095 with SE=0.003; negative values due to forcing non-overlapping gene sets), suggesting that scDRS scores computed using non-overlapping genes are still able to capture similarity between traits. These results are reported in the *Heterogeneous subpopulations of T cells…* subsection of the Results section (p.6), citing a new Supplementary Table 19.

(c) Do quantitative gene weights help capture differences between diseases/traits?

The scDRS disease scores are highly consistent between using vs. not using the quantitative gene weights (median correlation 0.98 across 74 traits; see response to Reviewer #2 Comment 1). Therefore, we do not believe that the use of quantitative gene weights substantially impacts differences between traits.

3. Since scDRS uses single cell level expression rather than summarizing by the cell type labels, the distribution of cell types in the dataset can bias the mean expression of the genes. For example, for a brain scRNA-seq dataset, neuronal cells might be dominant and the average expression of a gene is basically the average expression of neuronal cells rather than cells in the brain. The authors should discuss potential issues regarding the unbalanced dataset.

The reviewer makes a good point that scDRS computes gene-level statistics across all cells in the data set, and mean expression may thus be biased towards cell types or cell groups with disproportionately more cells in unbalanced datasets. We sought to assess the impact of this potential bias on the results reported in our manuscript by implementing a new option of scDRS that additionally takes a set of cell type annotations (or any cell group annotations) and adjusts for cell type proportions by inversely weighting cells by the number of cells in the corresponding cell type. We determined that the new option produced highly consistent results on the TMS FACS data (adjusting for proportions based on TMS FACS cell type labels; median score correlation 0.97 across 74 traits) and was well-calibrated in the null simulations (Supplementary Figure 4). We have updated the *Results across 120 TMS cell types…* subsection of the Results section (p.5), the Methods section (p.13), and the Supplementary Note (p.s5) to clarify this point.

We recommend the use of this new option only in the case of extremely unbalanced data sets, for 3 reasons. First, it produces consistent results for relatively balanced data sets such as TMS FACS. Second, it requires cell group annotations where the cell groups have a similar level of granularity (e.g., B cells vs. T cells instead of B cells vs. a subtype of CD4+ T helper 17 cells), which is not always available. For example, the TMS cell type annotation contains both high-level cell types like T cells and more fine-grained cell types like regulatory T cells. Third, the cell group annotation can be defined with different levels of granularity, such as broader types like immune cells or very specific types like CD4+ T helper 17 cells, and it is unclear how to choose the right level of granularity for a given data set. We have updated the Discussion section (p.9) and the Methods section (p.13) to discuss these points.

4. There is no mention about the computational cost and expected run time of this method. Is it scalable with the number of cells in the dataset? I see the largest dataset used in this study has 500K cells, but can you run this method for > 1 million cells in a reasonable time? For example, Tabula Sapiens has 2 million cells in the dataset.

scDRS scales linearly with the number of cells and the number of control gene sets (default 1,000) for both computation time and memory usage. We performed new benchmark experiments by subsampling cells from the Nathan et al. 500K data (Nathan et al. 2021 Nat Immunol, ref. 64) and, as expected, observed a linear relationship between the number of cells and both the computation time and memory usage (Supplementary Figure 3). Based on this result, it is estimated to take around 3 hours and 60GB of memory to run scDRS under the default setting (1,000 control gene sets) on a data set with *a million cells* and a similar level of sparsity as the Nathan et al. data, which is reasonable for high-performance computing clusters.

Of note, in this experiment, the memory usage (30G for the Nathan et al. 500K data) is only 1.5X of the theoretical lower limit, namely 18.9G consisting of 11.4G for loading the data in high precision (64-bit float) and 7.5G for computing the 1,000 sets of raw and normalized control scores for each cell (2*500,089*1000*8B=7.5G); the memory usage is 3X of the theoretical lower limit for low-precision computation. Therefore, we believe scDRS is reasonably efficient in memory usage. Based on this benchmark experiment, we also suggest an empirical formula for estimating the memory usage as 3*(low_precision_data_size + n_cell*n_control*8/1024^3).

We have updated the Methods section (p.12, citing a new Supplementary Figure 3) to report these findings and discuss these points. We have also updated the *Overview of methods* subsection of the Results section (p.3) to note that details on computational cost are provided in the Methods section.

5. I do think the authors have done great amount of work and performed comprehensive analyses but I found that the manuscript is a little hard to follow. I understand that there are many results to show but, especially for such high impact journal with broad readers, the authors should spend little more effort to make the main objectives and findings clearer. In addition, there are many repetitive expressions, which does not make it easier to read. It might be better to focus on main findings in the main text and move some of the analyses into supplementary information. For example, line 210-231 can be summarized in a few sentences and move details to supplementary information. In the 3 sections describing results of heterogeneous subpopulations of associated cells for autoimmune, brain-related and metabolic traits, the secondary analyses can be replaced with a few sentences and move details to supplementary information.

We have moved the details to the Supplementary Note for secondary analyses in the *Simulations assessing calibration and power* subsection of the Results section, the *Results across 120 TMS cell types …* subsection of the Results section, the *Heterogeneous subpopulations of T cells…* subsection of the Results section, and the *Heterogeneous subpopulations of hepatocytes…* subsection of the Results section. We also moved 6 of the 10 limitations / future direction points from the Discussion section to the Supplementary Note. Accordingly, we have reduced the length of the manuscript from 8,200 words to 6,800 words (17% shorter; excluding Methods and Figure captions) despite the addition of new content to address reviewer comments. We are open to further shortening the paper after all reviewer concerns have been confirmed to be completely addressed.

We have also identified and removed repetitive expressions in the paper. Specifically, we reduced the use of the terms “individual cells”, “individual cell-trait associations”, “associated cell type-disease pairs”, “MC test”, and “results are reported” appearing multiple times within a sentence. We also revised extremely long sentences used multiple times in the paper, such as “We verified that the inferred … score obtained by applying this procedure to independent … data were significantly correlated with …” used in the *Heterogeneous subpopulations of neurons …* subsection of the Results section (p.7) and the *Heterogeneous subpopulations of hepatocytes…* subsections of the Results section (p.7).

Decision Letter, first revision:

29th Mar 2022

Dear Dr Zhang,

Your Technical Report, "Polygenic enrichment distinguishes disease associations of individual cells in single-cell RNA-seq data" has now been seen by 2 referees. You will see from their comments below that while they find your work of interest, some important points are raised. We are interested in the possibility of publishing your study in Nature Genetics, but would like to consider your response to these concerns in the form of a revised manuscript before we make a final decision on publication.

To guide the scope of the revisions, the editors discuss the referee reports in detail within the team, 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 Technical Report format instructions, available
<a href="http://www.nature.com/ng/authors/article_types/index.html">here</a>.
Refer also to any guidelines provided in this letter.

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

A revised checklist is essential for re-review of the paper.

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

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

[REDACTED]

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

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

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

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

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

Sincerely,

Michael Fletcher, PhD
Associate Editor, Nature Genetics

ORCID: 0000-0003-1589-7087

Referee expertise:

Referee #1:

Referee #2:

Referee #3:

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
Report
The authors have addressed all the points raised.

Additional comments
1.Could the height result in Figure 5B be a reflection of the genetic correlation with ECOL (ie condition height GWAS results on ECOL and repeat?).
2.For Supp T14 why not list all the results from the software why just the 3 columns prop-sig-cell, ct-assoc-fdr, heterogeneity-fdr. It would be more transparent for the table to reflect the actual output of the software. As noted by the authors the minimum p-value from the 1000 cells is 1/1001 leading intern to a minimum ct-assoc-fdr value. As GWAS sample sizes get well-powered the number of significant cell-types can increase, with many listed at the minimum ct-assoc-fdr providing no information to discriminate between them. In contrast the assoc_mcz does does give indication of which cell-types may be more important. By not listing the output in the Supp Tables it suggests that this information is considered not useful. Yet, as demonstrated by the authors there is good agreement between the p-value from assoc_mcz and the assoc_mcp, but with some exceptions – presumably when the set of 1000 genes has some particular property through its size and LD distribution which is controlled for in the sampling of control genes. A pragmatic solution is to use the control set p-values to identify these outliers and then use test-statistic p-value to help prioritise the cell-types of relevance which could direct iPSC studies.

Reviewer #2:
Remarks to the Author:
The authors sufficiently answered my previous comments. I have no further concerns.

Author Rebuttal, first revision:

Response to reviewers for NG-TR58415R1

Reviewer #1

Thank you for the additional comments. We provide a point-by-point response below.

Additional comments
1. Could the height result in Figure 5B be a reflection of the genetic correlation with ECOL (ie condition height GWAS results on ECOL and repeat?).

We agree that the weak association between height and the proximal score of CA1 pyramidal neurons in Fig. 5B (P=0.006) may reflect the genetic correlation between height and ECOL (genetic correlation of 0.16, SE=0.01 via cross-trait LDSC). To investigate this, we performed a new conditional analysis by running scDRS on the top 1,000 height genes selected from the set of all genes excluding the top 1,000 ECOL genes; we performed the conditional analysis at the gene set level because GWAS results impact the scDRS results via the set of disease-relevant genes. We determined that the association between height and the proximal score of CA1 pyramidal neurons became non-significant in the conditional analysis (P=0.035, non-significant after Bonferroni correction for 8 traits tested (7 brain-related traits and height)), suggesting that the slight association between height and the proximal score of CA1 pyramidal neurons may reflect the genetic correlation between height and ECOL. We have updated the *Heterogeneous subpopulations of neurons…* subsection of the Results section (p.7), citing new content in Supp. Table 25, to include this result. (However, we elected to continue to report height results in Fig. 5B, for consistency with Fig. 4E and Fig. 5D.)

2. For Supp T14 why not list all the results from the software why just the 3 columns prop-sig-cell, ct-assoc-fdr, heterogeneity-fdr. It would be more transparent for the table to reflect the actual output of the software. As noted by the authors the minimum p-value from the 1000 cells is 1/1001 leading intern to a minimum ct-assoc-fdr value. As GWAS sample sizes get well-powered the number of significant cell-types can increase, with many listed at the minimum ct-assoc-fdr providing no information to discriminate between them. In contrast the assoc_mcz does does give indication of which cell-types may be more important. By not listing the output in the Supp Tables it suggests that this information is considered not useful. Yet, as demonstrated by the authors there is good agreement between the p-value from assoc_mcz and the assoc_mcp, but with some exceptions – presumably when the set of 1000 genes has some particular property through its size and LD distribution which is controlled for in the sampling of control genes. A pragmatic solution is to use the control set p-values to identify these outliers and then use test-statistic p-value to help prioritise the cell-types of relevance which could direct iPSC studies.

As suggested by the reviewer, we now include MC control set p-values (column labels “assoc-mc-pval” and “hetero-mc-pval”) and MC test-statistic z-scores (column labels “assoc-mc-zsc” and “hetero-mc-zsc”) for both cell type-disease association and within-cell type disease association heterogeneity in Supp. Tables 14-16,24 (in addition to prop-sig-cell, ct-assoc-fdr, and heterogeneity-fdr). We agree that the MC test-statistic z-scores can be further used to prioritize cell types whose MC control set p-values have reached the limit of 1/1001, and have updated the Methods section to recommend this use (p.11).

We also note that the full results (disease + control scores for the 74 traits on TMS FACS) are available online (https://figshare.com/articles/dataset/scDRS_data_release_030122_score_file_tmsfacs/19312607), as noted in the Data availability section (p.16).

Decision Letter, second revision:

Our ref: NG-TR58415R2

12th Apr 2022

Dear Martin,

Thank you for submitting your revised manuscript "Polygenic enrichment distinguishes disease associations of individual cells in single-cell RNA-seq data" (NG-TR58415R2).

We have made an editorial check of your response to the last round of comments and we think they are satisfactory and do not require further review. Therefore, we'll be happy in principle to publish your manuscript in Nature Genetics, pending minor revisions 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)-- 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

Final Decision Letter:

In reply please quote: NG-TR58415R3 Zhang

19th Jul 2022

Dear Martin,

I am delighted to say that your manuscript "Polygenic enrichment distinguishes disease associations of individual cells in single-cell RNA-seq data" has been accepted for publication in an upcoming issue of Nature Genetics.

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

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

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

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

Your paper will be published online after we receive your corrections and will appear in print in the next available issue. You can find out your date of online publication by contacting the Nature Press Office (press@nature.com) after sending your e-proof corrections. Now is the time to inform your Public Relations or Press Office about your paper, as they might be interested in promoting its publication. This will allow them time to prepare an accurate and satisfactory press release. Include your manuscript tracking number (NG-TR58415R3) and the name of the journal, which they will need when they contact our Press Office.

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

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

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

Authors may need to take specific actions to achieve <a href="https://www.springernature.com/gp/open-research/funding/policy-compliance-

faqs"> compliance</a> with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to <a href="https://www.springernature.com/gp/open-research/plan-s-compliance">Plan S principles</a>) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, 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 first submitted after 1 January, 2021.

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

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

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

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

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

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

If you have not already done so, we invite you to upload the step-by-step protocols used in this manuscript to the Protocols Exchange, part of our on-line web resource, natureprotocols.com. If you complete the upload by the time you receive your manuscript proofs, we can insert links in your article that lead directly to the protocol details. Your protocol will be made freely available upon publication of your paper. By participating in natureprotocols.com, you are enabling researchers to more readily reproduce or adapt the methodology you use. Natureprotocols.com is fully searchable, providing your protocols and paper with increased utility and visibility. Please submit your protocol to https://protocolexchange.researchsquare.com/. After entering your nature.com username and password you will need to enter your manuscript number (NG-TR58415R3). 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