nature research

Peer Review Information

Journal: Nature Methods
Manuscript Title: Enhanced recovery of single-cell RNA-sequencing reads for missing gene expression data
Corresponding author name(s): Allan-Hermann Pool, Yuki Oka

Editorial Notes:

Reviewer Comments & Decisions:

Decision Letter, initial version:

1st Aug 2021

Dear Dr Oka,

Thank you for your inquiry about submitting your manuscript, "Enhanced recovery of single-cell RNA-sequencing reads for missing gene expression data" to Nature Methods. The paper sounds like it may be of interest, and should fit the scope of the journal. We would be willing to read the full manuscript.

Of course, it is very difficult to judge a paper based only on the limited information available in a presubmission inquiry. Therefore I am sure you understand that we cannot promise to send your paper out for peer review and must read it in its entirety before deciding if this would be suitable.

Please keep in mind that the journal is aimed at a large, interdisciplinary audience and places a strong emphasis on the practical value of the work presented. We strongly encourage you to include data to validate method performance and demonstrate its general applicability.

Manuscripts describing new algorithms and software should describe the principles underlying the algorithms in an accessible style targeted for our broad biological audience. The manuscript should demonstrate the practical relevance of the tool and any advantages and limitations it presents compared to available software. The code, and ideally a user manual and example data for testing the code, must be made available to reviewers as part of the peer review process.

You will find our Guide to Authors at http://www.nature.com/naturemethods to assist you in preparing your manuscript. However, it is not necessary at this stage to spend major effort adhering to our

detailed formatting instructions.

Thank you for your interest in Nature Methods.

Sincerely,

Lin Tang, PhD
Senior Editor
Nature Methods

Guided Open Access:
If you are interested in publishing your paper open access, we encourage you to utilize our Guided Open Access system, in which primary research papers are simultaneously considered for publication at Nature Methods, Nature Communications, and Communications Biology. Authors whose papers meet the editorial bar for potential publication in at least one of these journals will receive a structured Editorial Assessment Report, including peer reviews and clear and constructive recommendations for revising for one or more of these journals. Guided Open Access can also remove the potential need for sequential submissions (should Nature Methods decline to review your paper), allowing for a more rapid and efficient editorial assessment. For articles accepted into Nature Methods, the APC (including both the initial Editorial Assessment Charge and the top-up APC payment) is ~50% less than the direct APC. More information about Guided Open Access can be found here:
https://www.nature.com/nature-research/open-access/guided-open-access;
https://www.nature.com/articles/s41592-021-01073-y.

Author Rebuttal to Initial comments

[There is no rebuttal letter at this stage.]

Decision Letter, first revision:

26th Apr 2022

Dear Dr Oka,

Thank you for submitting your manuscript entitled "Enhanced recovery of single-cell RNA-sequencing reads for missing gene expression data". We have given the paper our careful consideration but we regret that we cannot publish it in Nature Methods.

It is Nature Methods' policy to decline a substantial proportion of manuscripts without peer-review, so that they may be sent elsewhere without delay. Decisions of this kind are made by the editorial staff when it appears that papers are unlikely to succeed in the competition for limited space.

Among the considerations that arise at this stage are a manuscript's probable interest, level of methodological development and immediate practical relevance to a general readership. We do not doubt the technical quality of your work in analyzing read loss and developing methods to improve analysis of Droplet based single-cell RNA-seq data, or that it will be of interest to others working in this area of research. However, I am sorry to say we do not think that the technical advances presented will have a sufficiently significant impact on a broader readership to justify publication in Nature Methods.

Although we cannot offer to publish your manuscript, we recommend that you consider transferring your manuscript to our sister journal, Communications Biology, a selective open-access Nature Research title that publishes research bringing new insight into a focused area of biology (www.nature.com/commsbio). For more information, please see the footnote below, which includes a link you can use to transfer your manuscript.

Nevertheless, thank you very much for giving us the opportunity to consider your manuscript. I am sorry that we cannot be more positive on this occasion and hope that you will promptly find a more appropriate forum for presenting your work.

Sincerely,

Lin Tang, PhD
Senior Editor
Nature Methods

**Your manuscript has been recommended for Communications Biology based on our familiarity with the journal's criteria. While our editorial teams are independent and the editors there will make their own decision, recommended manuscripts have a higher probability of success. Manuscripts submitted to Communications Biology receive their first editorial decision before peer review in an average of 9 days. If you have any questions about Communications Biology, please contact the Chief Editor, Dr Brooke LaFlamme (brooke.laflamme@us.nature.com). Please note that Communications Biology is a fully open-access journal and an article processing charge will apply to any papers accepted for publication. Our open access pages contain information about article processing charges, open access funding, and advice and support from Springer Nature.

[REDACTED]

If you transfer to Nature journals or to the Communications journals, you will not have to re-supply manuscript metadata and files, unless you wish to make modifications. This link can only be used once and remains active until used.

All Nature Portfolio journals are editorially independent, and the decision on your manuscript will be taken by their editors. For more information, please see our manuscript transfer FAQ page.

Note that any decision to opt in to In Review at the original journal is not sent to the receiving journal

on transfer. You can opt in to In Review at receiving journals that support this service by choosing to modify your manuscript on transfer. In Review is available for primary research manuscript types only.

** For Nature Research Group general information and news for authors, see http://npg.nature.com/authors

Author Rebuttal, first revision:

[There is no rebuttal letter at this stage.]

Decision Letter, second revision:

20th Sep 2022

Dear Dr Oka,

Thank you very much for your patience during the peer review of your manuscript entitled "Enhanced recovery of single-cell RNA-sequencing reads for missing gene expression data". The paper has now been seen by 2 reviewers, whose comments are attached. While they find your work of potential interest, they have raised serious concerns which in our view are sufficiently important that they preclude publication of the work in Nature Methods, at least in its present form.

As you will see, the reviewers raise concerns about different aspects of the study. Should substantial revisions (including additional analysis and improvement to the software) allow you to fully address these criticisms we would be willing to look at a revised manuscript (unless, of course, something similar has by then been accepted at Nature Methods or appeared elsewhere). This includes submission or publication of a portion of this work somewhere else. We hope you understand that until we have read the revised paper in its entirety we cannot promise that it will be sent back for peer-review.

If you are interested in revising this manuscript for submission to Nature Methods in the future, please contact me to discuss your appeal before making any revisions. Otherwise, we hope that you find the reviewers' comments helpful when preparing your paper for submission elsewhere.

Thank you for the opportunity to consider your work.

Sincerely,

Lin Tang, PhD
Senior Editor
Nature Methods

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
The manuscript by Pool et al. makes the case for refining the gene annotations used in scRNA-seq analysis to capture more signal in data generated using 3' protocols.

Currently, various read processing methods allow the inclusion of intron reads by default or as an option in the gene counting step (e.g. zUMIs https://pubmed.ncbi.nlm.nih.gov/29846586/, alevin-fry (spliced + intron reference, https://www.nature.com/articles/s41592-022-01408-3) - please consider citing these works), CellRanger v7 https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/release-notes). However, reads lost to 3' UTR annotation issues or overlapping features has not been systematically explored to my knowledge so bringing this altogether to look at their relative contribution to signal is valuable.

General questions:

How tissue specific are these results i.e. does the relative contribution of signal from the 3 sources vary depending upon whether you are looking at brain, blood vs other tissues? This study uses mouse brain tissue and a PBMC sample, which is a good start, but given the large amount of public scRNA-seq data available, it should be relatively easy to explore this in other sample / cell types (perhaps an atlas dataset?) to assess how generalisable these findings are.

Does the optimized annotation have to be redefined for every experiment / different scRNA-seq protocols, or is it generalisable? How feasible is the case-by-case scrutiny for recovering intergenic reads from 3' unannotated exons and how long does this step take? If these signals are highly dataset specific (i.e. only relevant for brain mouse samples and PBMC human samples) and require significant optimization, this may limit the usefulness of the online resources provided by the authors.

Of the interventions suggested, it appears that intron inclusion boosts signal the most (Fig 1e and g), although many new genes are added through both intron or 3' intergenic mapping (Fig 1f). I wondered about the relative contribution of these new genes in downstream analysis though. Are many of the newly detected genes highly variable and drivers of the new clustering results, or are they mostly lowly expressed / uninformative and likely to be filtered out from downstream analysis as part of quality control? Or perhaps the two analyses (exonic reference vs optimized reference) are based on signal from mostly the same genes, just with enhanced signal that has improved cell type specificity in the reference optimized results? It would be good to do some analysis to tease this apart to provide the reader with further insights on how many of the 'missed' features are likely to turn up as genes of interest (hyper-variable / marker genes). A number of examples are provided in Fig 5 and Extended Fig 3, but it wasn't clear if these were it, or if there are any more such examples.

What are the signal / biotype characteristics of the unique genes in Extended Fig 2 i.e. are they lowly expressed features, or do they span the intensity range? Are particular gene biotypes over-represented in this category?

Specific comments:

Consider combining Figures 1 and 2?
Consider providing a CHM13 version of the human annotation?
I'm not sure that the claim in the sentence that begins on line 292 with 'These approaches…' is correct. It was my understanding that imputation methods are often applied to 3' scRNA-seq data.
Typos on line 298, 424, 450, 477
Add more detail to Methods on line 674 (e.g. what options were used for filtering poor quality cells / genes with low abundance / normalization / clustering in Seurat? Given the results in Figure 5, the methods here seem incomplete).

Reviewer #2:
Remarks to the Author:
**** A. Summary of the key results

This paper introduced a workflow for the construction of an optimized transcriptomic reference for improving the profiling resolution (specifically, increasing the sensitivity of detected genes) of single-cell RNA-sequencing data analysis. In the proposed workflow, to build the optimized reference transcriptome, the following steps are taken:

1. The exonic reference is replaced by a pre-mRNA reference
2. Overlapping genes are removed if satisfying certain criteria
3. The 3' UTR of qualified genes are extended

The key results are:
Quantitative evaluation of the three main mechanisms that cause loss of reads in the traditional exonic-reference-based analysis.
Demonstration of a general strategy to build an optimized reference transcriptome.
Description and evaluation of the benefits of replacing the traditional exonic reference with the optimized reference transcriptome in single-cell data analysis. Specifically, the demonstration that thousands of genes absent in the exonic-reference-based analysis can be recovered, and new cell types are identified by those newly recovered genes in both a human and a mouse dataset.

**** B. Originality and significance: if not novel, please include reference

The key approaches introduced in this manuscript extend, expand upon and refine a previous paper published by the same first author, "Pool, A.-H. et al. The cellular basis of distinct thirst modalities. Nature (2020). " In the previous paper, the expanded transcriptome is built by including unspliced transcripts, removing 294 pseudogenes with little to no exonic read mapping from the genome, and extending 18 genes' untranslated regions. In the current manuscript, the construction process of the expanded transcriptome is expanded and formalized into a much more general computational workflow. The authors implemented the workflow in R so that users can construct the optimized transcriptome in a partially automated way (though with the help of some manual inspection). The authors also provided an optimized transcriptome reference for mouse and human.

Part of the proposed workflow, specifically, the inclusion of intronic reads in the reference, has been adopted by some existing quantification tools. For example, Cell Ranger 7 includes the intronic reads

in the quantification process by default (see https://support.10xgenomics.com/docs/intron-mode-rec?src=social&lss=linkedin&cnm=soc-li-ra_g-program&cid=7011P000000y072). Starsolo (Kaminow et al., bioRxiv, 2021) has the option `--soloFeatures GeneFull` to include the intronic reads in its procedure, as described in section 5.1.3 in starsolo preprint.
The splici (spliced + intronic) transcriptomic reference used in alevin-fry (He et al., Nature Methods, 2022) contains the introns of genes, as demonstrated in Supplementary section S2 in its manuscript. Kallisto|bustools (Melsted et al., Nat Biotech, 2021) can take into account intronic reads if making corresponding changes and passing `--workflow lamanno` in its pipeline.

On top of these existing methods, the significance of this manuscript comes from the fact that this optimized reference transcriptome further reduced the loss of reads, compared with the existing tools, by removing low-quality overlapping genes and extending the 3' UTR of genes that are considered as having unannotated 3' UTR region. These two strategies have rarely been explored and have the potential to be adopted as a standard step for single-cell data processing pipeline. Moreover, the author provided the R scripts for building the optimized reference transcriptome for any given genome.

**** C. Data & methodology: validity of the approach, quality of data, quality of presentation

The Methods section was well written. It is easy to follow the text and the corresponding figures to understand the process of building an optimized reference transcriptome. The only exception is that the algorithm in the "Resolving gene overlap derived read loss" section is a bit hard to follow. It would be great if a figure plus a formal algorithmic style description (e.g. more in the style of pseudocode) could be added to help demonstrate the process. Additionally, I have the following concerns.

The GitHub repository provided in the manuscript, https://github.com/PoolLab/Generecovery, doesn't contain all the code of the steps described in the manuscript. For example, in the file 1_Annotation pre-processing.R, step A: resolve gene overlap derived read loss only contains the code for step A1: Generate a rank-ordered gene list of same-strand overlapping genes, but not A2: Prioritize gene list by classifying overlapping into recommended action categories. Moreover, In the provided R script, bash commands also exist, and some strings (gene names and paths) are hard-coded in the script, making the workflow hard to apply by the user and near impossible to reproduce precisely. At the very least, any dependent or necessary files should be included in the repository and referenced by an appropriate relative path so that the scripts can easily be adopted by a user. Please see section F below for a more detailed suggestion.

The mouse and human datasets included in the paper clearly show that the three problems of the current reference are outstanding and need to be addressed. Figure 5 shows that the optimized reference transcriptome is able to recover thousands of genes and therefore improve the cell type identification result. One concern is that the code used for generating figure 5 is missing in the GitHub repo. Meanwhile, in the Methods section, the process of clustering and labeling the cells is unclear. For example, the parameters used in the clustering analysis, such as the number of variable features and PCs used for initial dimensionality reduction, the number of nearest neighbors for computing k-NN graph, and the resolution for the clustering algorithm (if following the standard Seurat pipeline, https://satijalab.org/seurat/articles/pbmc3k_tutorial.html), are not specified.

**** D. Appropriate use of statistics and treatment of uncertainties

These were used appropriately.

**** E. Conclusions: robustness, validity, reliability

The section is clear and easy to follow. It demonstrates the problem that this work wants to address and the contributions this work makes.

**** F. Suggested improvements: experiments, data for possible revision

It is not clear from the paper how much of the gains of replacing the exonic transcriptomic reference with the optimized reference transcriptome in the downstream analyses are from the inclusion of the introns, which has been adopted by several other methods, and how much additional is gained from removing overlapping genes and extending 3' UTR of genes. If possible, a follow-up experiment can be added to show the importance of the newly recovered genes by addressing each of the three problems for the potential downstream analyses. For the genes recovered by addressing each of the three problems, how many of them are protein-coding genes? How many of them participate in important pathways of the studied samples? How many of them are the cell type markers of the studied samples? Are those genes significantly differentially expressed in one of the newly discovered clusters? Are all of these additional changes necessary to discover the new cell types / clusters, or do just introns or just extended 3' UTRs suffice? The purpose of asking this follow-up experiment is to show the importance of addressing each of the three problems for the proposed improvement in the paper.

Another suggestion is that the existing pipelines like STARsolo have the ability to include intronic reads in the final cellular gene expression data by specifying some flags and the ability to solve the multimapping reads by an EM framework. The inclusion of the intronic reads corresponds to the inclusion of the unspliced transcripts in the current manuscript. Likewise, the EM workflow is designed to solve the loss of reads caused by overlapping genes because, instead of discarding the multimapping reads, the EM algorithm will try to appropriately assign UMIs probabilistically based on all the observed data (including the multimapped reads) in each cell. Therefore, it would be useful if the authors could compare the results from the optimized reference transcriptome with those from the existing pipelines by setting the flag of including intronic reads and multimapping reads. For instance, in STARsolo, the user should specify --soloFeatures GeneFull and --soloMultiMappers EM. A very direct and meaningful comparison would be to show e.g. the effect of just removing the overlapping genes versus resolving the multimapping reads using the EM algorithm — is removing the overlappers necessary? Does it perform better than STARsolo's probabilistic resolution?
Figure 5 is informative, clean, and quite striking. However, it is hard to understand the result without the code or a detailed description of the analysis in the manuscript. To be specific, it is not clear if the new clusters (not cell types, the clusters returned by the Seurat `FindClusters()` function) can be detected by increasing the resolution in the exonic-reference-based analysis; it is not clear if the cells in each cluster after expanding the reference are the same as before; it is not clear if the newly discovered genes can be detected as the highly variable genes by the Seurat `FindMarkers()` function to be used for identifying cell types in the standard Seurat pipeline (https://satijalab.org/seurat/articles/pbmc3k_tutorial.html). It is unclear whether the newly discovered genes are the only markers of the newly discovered cell types. I would like to suggest the

authors upload an R or Python notebook of the analysis of generating Figure 5 and have an explanation of the process of cell type assignment in the Methods section.
A follow-up suggestion is that because the marker genes included in Figure 5 are very important, it will be useful to show where the reads of those genes map and by which mechanism (intronic reads, unannotated 3' UTR, or overlap genes) those marker genes failed to be recovered by the exonic reference.

The optimized reference transcriptome has the potential to become a key step in standard single-cell analysis. However, according to the codebase in the provided GitHub repository, the current implementation can't be directly utilized by the community in a straightforward and generalizable manner. The usability of the method itself is likely to be quite important to its adoption. I strongly suggest the authors develop a properly-documented and easily-installable self-contained package for the construction of the optimized reference transcriptome. Ideally, this package could be submitted to Bioconductor or CRAN, or if implemented in python, could be distributed via pip or bioconda. At the very least, the code should be made into an R package, with all required dependencies, installable from its GitHub repository via `devtools` and proper documentation and vignette should be provided. Additionally, some tutorials and example runs demonstrate using the package to generate and quantify against the optimized reference transcriptome would be helpful to potential users, though such additional resources aren't strictly necessary if a full package with documentation is available.

**** G. References: appropriate credit to previous work?

The authors gave credit to almost all previous work. One small suggestion is that the author could mention that many existing methods that have the ability to process intronic reads, like CellRanger, kallisto|bustools, STARsolo, zUMIs, alevin, and alevin-fry.

**** H. Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction, and conclusions

Overall, the manuscript is well written and easy to follow. The authors clearly expressed the rationale of the work, the details of the implementation, and the contribution and significance of the proposed optimized reference transcriptome. I believe the authors have proposed some straightforward and concrete improvements to the reference transcriptome curation process that has the potential to positively impact a broad variety of single-cell sequencing analysis studies.

Additionally, the typos I noticed while reviewing the manuscript can be found below.

10X and 10x are used throughout the manuscript. It should be 10x Genomics.

scRNA-seq, scRNA-sequencing, and single-cell are all used.

Read-through, readthrough and read through are used interchangeably.

Page1
Line 33: three-step

Page 4

Line 75: remove (iv)

Page 8:
Line 178-179: “premature start transcripts”. The quotes are of wrong format.

Page 19:
Line 429: Read itentity -> read identity

Page 20:
Line 449: missing space after closing parenthesis

Page 23:
Line 506: were and 10 characters. Delete and
Line 507: reads was -> reads were
Line 508: indeces -> indices
Line 510: to generating -> to generate

Page 29:
Line 641: the latter -> the former (if I understood correctly)

Page 30:
Line 663: to e new -> to a new

Reviewer #3:
None

Although we cannot publish your paper, it may be appropriate for another journal in the Nature Portfolio. If you wish to explore the journals and transfer your manuscript please use [REDACTED] manuscript transfer FAQ page.

Note that any decision to opt in to In Review at the original journal is not sent to the receiving journal on transfer. You can opt in to In Review at receiving journals that support this service by choosing to modify your manuscript on transfer. In Review is available for primary research manuscript types only.

Author Rebuttal, second revision:

Plan of revisions for Pool et al. 2022 submission to Nature methods

We appreciate the reviewers' positive and constructive suggestions. Please see our revision plans in blue.

Reviewer 1:

The manuscript by Pool et al. makes the case for refining the gene annotationsused in scRNA-seq analysis to capture more signal in data generated using 3'protocols. Currently, various read processing methods allow the inclusion of intron reads by default or as an option in the gene counting step (e.g. zUMIshttps://pubmed.ncbi.nlm.nih.gov/29846586/, alevin-fry (spliced + intron reference,https://www.nature.com/articles/s41592-022-01408-3) - please consider citing these works), CellRanger v7 https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/release-notes).

Thank you, we will add these references. Note however, that the different strategies and software for integrating intronic reads vary significantly (detection of hundreds of genes and significant expression level differences for thousands of genes). We have characterized that in supplemental figure 2 and will extend that figure to compare existing methods with the optimized reference.

However, reads lost to 3' UTRannotation issues or overlapping features has not been systematically explored to my knowledge so bringing this altogether to look at their relative contribution to signal is valuable.

General questions: How tissue specific are these results i.e. does the relative contribution of signal from the 3 sources vary depending upon whether you are looking at brain, blood vs other tissues? This study uses mouse brain tissue and a PBMC sample, which is a good start, but given the large amount of public scRNA-seq data available, it should be

relatively easy to explore this in other sample / cell types (perhaps an atlas dataset?) to assess how generalisable these findings are.

This is an important suggestion. We will evaluate the performance of the optimized references in half a dozen distinct mouse and human tissues (thymus, blood, tongue, heart, lung, pancreas) from the *Tabula muris* (https://www.nature.com/articles/s41586-018-0590-4) and *Tabula sapiens* (https://www.science.org/doi/10.1126/science.abl4896) consortium datasets, respectively.

Does the optimized annotation have to be redefined for every experiment /different scRNA-seq protocols, or is it generalisable?

Resolving gene overlaps needs to be done only once per genome sequence and can be used for any dataset. The intergenic annotations will likely benefit from iterative improvements with different datasets as different genes are active in distinct tissues. We will directly evaluate that on Tabula muris and Tabula sapiens datasets to quantify that.

How feasible is the case-by-case scrutiny for recovering intergenic reads from 3' unannotated exons and how long does this step take? If these signals are highly dataset specific (i.e. only relevant for brain mouse samples and PBMC human samples) and require significant optimization, this may limit the usefulness of the online resources provided by the authors.

Case by case scrutiny and quantification is highly feasible and we plan to do that in the revision. Once the first pass is done (which we have done for mouse and human) the remaining iterations are fast. One tangible benefit for our optimized references is that we have already performed the time-consuming portions of this work, which make further improvements efficient and fast.

Of the interventions suggested, it appears that intron inclusion boosts signal the most (Fig 1e and g), although many new genes are added through both intron or 3'intergenic mapping (Fig 1f). I wondered about the relative contribution of these new genes in downstream analysis though. Are many of the newly detected genes highly variable and drivers of the new clustering results, or are they mostly lowly expressed / uninformative and likely to be filtered out from downstream analysis as part of quality control? Or perhaps the two analyses (exonic reference vs optimizedreference) are based on signal from mostly the same genes, just with enhanced signal that has improved cell type specificity in the reference optimized results? It would be good to do some analysis to tease this apart to provide the reader with further insights on how many of the 'missed'

features are likely to turn up as genes of interest (hyper-variable / marker genes). A number of examples are provided in Fig 5 and Extended Fig 3, but it wasn't clear if these were it, or if there are any more such examples.

In general, using the optimized references significantly changes the composition of the most variable genes (~30-40%) and dramatically improves the ability to tell apart distinct cell types and states (currently addressed in Figure 5). We can directly show the contribution of each of these improvements (resolving gene overlaps, including intronic reads and adding unannotated 3'UTRs) and add this to Figure 5. We can also evaluate how much these approaches vary by distinct tissue type by evaluating that in *Tabula muris* and *Tabula sapiens* datasets, which we will add as a supplemental figure.

The annotation problems impact gene categories uniformly (growth factors, membrane proteins, transcription factors etc.). We will do a GO term enrichment analysis for genes impacted by the different annotation issues and add as a supplemental figure to address this issue.

What are the signal / biotype characteristics of the unique genes in Extended Fig 2i.e. are they lowly expressed features, or do they span the intensity range? Are particular gene biotypes over-represented in this category?

There is a surprising expression range from high to low expressing genes that get differentially detected with different intronic read integration solutions. Furthermore, we found that for thousands of detected "shared" genes, the detection levels vary dramatically depending on which intronic inclusion approach is used. We will extend this analysis in Extended Fig 2 to quantify both of these issues.

Are particular gene biotypes over-represented in this category?

We will include GO term enrichment analysis for Extended data Fig 2.

Specific comments:

- Consider combining Figures 1 and 2?
- Consider providing a CHM13 version of the human annotation?

Will generate.

I'm not sure that the claim in the sentence that begins on line 292 with 'These approaches…' is correct. It was my understanding that imputation methods are often applied to 3' scRNA-seq data.

Imputation strategies that are based on full length transcript sequencing data would start to address some of the issues fixed by the optimized reference. Will specify in text.

Typos on line 298, 424, 450, 477
Add more detail to Methods on line 674 (e.g. what options were used for filtering poor quality cells / genes with low abundance / normalization / clustering in Seurat? Given the results in Figure 5, the methods here seem incomplete).
Will add in Methods section and also provide code in github.

Reviewer #2:

A. Summary of the key results

This paper introduced a workflow for the construction of an optimized transcriptomic reference for improving the profiling resolution (specifically, increasing the sensitivity of detected genes) of single-cell RNA-sequencing data analysis. In the proposed workflow, to build the optimized reference transcriptome, the following steps are taken:
1. The exonic reference is replaced by a pre-mRNA reference

Note: we come up with a hybrid intronic read capture solution that overcomes issues with other methods.

2. Overlapping genes are removed if satisfying certain criteria
3. The 3' UTR of qualified genes are extended

The key results are:
Quantitative evaluation of the three main mechanisms that cause loss of reads in the traditional exonic-reference-based analysis. Demonstration of a general strategy to build an optimized reference transcriptome. Description and evaluation of the benefits of replacing the traditional exonic reference with the optimized reference transcriptome in single-cell data analysis. Specifically, the demonstration that thousands of genes absent in the exonic-reference-based analysis can be recovered, and new cell types are identified by those newly recovered genes in both a human and a mouse dataset.

B. Originality and significance: if not novel, please include reference

The key approaches introduced in this manuscript extend, expand upon and refine a previous paper published by the same first author, "Pool, A.-H. et al. The cellularbasis of distinct thirst modalities. Nature (2020). " In the previous paper, the expanded

transcriptome is built by including unspliced transcripts, removing 294 pseudogenes with little to no exonic read mapping from the genome, and extending 18 genes' untranslated regions. In the current manuscript, the construction process of the expanded transcriptome is expanded and formalized into a much more general computational workflow. The authors implemented the workflow in R so that users can construct the optimized transcriptome in a partially automated way (though with the help of some manual inspection). The authors also provided an optimized transcriptome reference for mouse and human. Part of the proposed workflow, specifically, the inclusion of intronic reads in the reference, has been adopted by some existing quantification tools. For example, Cell Ranger 7 includes the intronic reads in the quantification process by default (see https://support.10xgenomics.com/docs/intron-mode-rec?src=social&lss=linkedin&cnm=soc-li-ra_g-program&cid=7011P000000y072). Starsolo (Kaminow et al., bioRxiv, 2021) has the option `--soloFeatures GeneFull` to include the intronic reads in its procedure, as described in section 5.1.3 in starsolo preprint. The splici (spliced + intronic) transcriptomic reference used in alevin-fry (He et al.,Nature Methods, 2022) contains the introns of genes, as demonstrated in Supplementary section S2 in its manuscript. Kallisto bustools (Melsted et al., Nat Biotech, 2021) can take into account intronic reads if making corresponding changes and passing `--workflow lamanno` in its pipeline.

An important takeaway of the paper is that the precise strategy for including intronic reads has a significant impact on which genes are detected and how efficiently (Extended data Figure 2). We will extend this analysis as proposed in the Reviewer 1 section to characterize how different strategies for including intronic reads influence gene detection and gene expression estimates.

On top of these existing methods, the significance of this manuscript comes from the fact that this optimized reference transcriptome further reduced the loss of reads, compared with the existing tools, by removing low-quality overlapping genes and extending the 3' UTR of genes that are considered as having unannotated 3' UTR region. These two strategies have rarely been explored and have the potential to be adopted as a standard step for single-cell data processing pipeline. Moreover, the author provided the R scripts for building the optimized reference transcriptome for any given genome.

We thank the reviewer for the enthusiastic comments.

C. Data & methodology: validity of the approach, quality of data, quality of presentation

The Methods section was well written. It is easy to follow the text and the corresponding figures to understand the process of building an optimized reference transcriptome. The

only exception is that the algorithm in the “Resolving gene overlap derived read loss” section is a bit hard to follow. It would be great if a figure plus a formal algorithmic style description (e.g. more in the style of pseudocode) could be added to help demonstrate the process.

We will add figure and provide pseudocode.

Additionally, I have the following concerns. The GitHub repository provided in the manuscript, https://github.com/PoolLab/Generecovery, doesn’t contain all the code of the steps described in the manuscript. For example, in the file 1_Annotation pre-processing.R, step A: resolve gene overlap derived read loss only contains the code for step A1:Generate a rank-ordered gene list of same-strand overlapping genes, but not A2: Prioritize gene list by classifying overlapping into recommended action categories.

We apologize for this inconvenience. We will add section A2 in R on Github.

Moreover, In the provided R script, bash commands also exist, and some strings (gene names and paths) are hard-coded in the script, making the workflow hard to apply by the user and near impossible to reproduce precisely. At the very least, any dependent or necessary files should be included in the repository and referenced by an appropriate relative path so that the scripts can easily be adopted by a user. Please see section F below for a more detailed suggestion.

Along the same line, we will add all dependent files to Github repository and will add relative links to them.

The mouse and human datasets included in the paper clearly show that the three problems of the current reference are outstanding and need to be addressed. Figure 5 shows that the optimized reference transcriptome is able to recover thousands of genes and therefore improve the cell type identification result. One concern is that the code used for generating figure 5 is missing in the GitHub repo. Meanwhile, in the Methods section, the process of clustering and labeling the cells is unclear. For example, the parameters used in the clustering analysis, such as the number of variable features and PCs used for initial dimensionality reduction, the number of nearest neighbors for computing k-NN graph, and the resolution for the clustering algorithm (if following the standard Seurat pipeline,https://satijalab.org/seurat/articles/pbmc3k_tutorial.html), are not specified.

The analysis parameters for the comparative analysis were identical (same number of variable genes, PC-s, clustering resolution etc.). We will both specify these details in the Methods section and add the code in Github with dependent files.

D. Appropriate use of statistics and treatment of uncertainties
These were used appropriately.

E. Conclusions: robustness, validity, reliability

The section is clear and easy to follow. It demonstrates the problem that this work wants to address and the contributions this work makes.

F. Suggested improvements: experiments, data for possible revision

It is not clear from the paper how much of the gains of replacing the exonic transcriptomic reference with the optimized reference transcriptome in the downstream analyses are from the inclusion of the introns, which has been adopted by several other methods, and how much additional is gained from removing overlapping genes and extending 3' UTR of genes. If possible, a follow-up experiment can be added to show the importance of the newly recovered genes by addressing each of the three problems for the potential downstream analyses.

We will analyze the relative contributions (detected genes and recovered reads) of fixing the three problems in Figure 5 and will quantify that for different datasets in Tabula muris and Tabula sapiens datasets in a new extended figure.

For the genes recovered by addressing each of the three problems, how many of them are protein-coding genes?

Will add quantification in supplemental figure.

How many of them participate in important pathways of the studied samples? How many of them are the cell type markers of the studied samples? Are those genes significantly differentially expressed in one of the newly discovered clusters?

All these questions are addressed in Figure 5 and in Extended Figure 3. However, for better presentation, we will expand analysis in Extended Figure 3 to quantify that for mouse brain and human PBMCs.

Are all of these additional changes necessary to discover the new cell types / clusters, or do just introns or just extended 3' UTRs suffice? The purpose of asking this follow-up experiment is to show the importance of addressing each of the three problems for the proposed improvement in the paper.

The majority of improvement comes from including intronic reads, followed by resolving gene overlap derived read loss and then integrating intergenic reads in unannotated 3' UTRs. However, the best result was obtained under the combined corrections.

Another suggestion is that the existing pipelines like STARsolo have the ability to include intronic reads in the final cellular gene expression data by specifying flags and the ability to solve the multimapping reads by an EM framework.

While including intronic reads can be achieved by several packages, we demonstrate that detection of hundreds of genes and significant expression level differences of thousands of genes depends on the precise strategy for including intronic reads (Extended data figure 2). We propose a hybrid strategy that outperforms STARsolo's and Cellranger's intronic integrating solutions. We will compare those head-to-head and extend this analysis in Extended Figure 2 to characterize and quantify this often ignored problem of intronic read integration.

The inclusion of the intronic reads corresponds to the inclusion of the unspliced transcripts in the current manuscript. Likewise, the EM workflow is designed to solve the loss of reads caused by overlapping genes because, instead of discarding the multimapping reads, the EM algorithm will try to appropriately assign UMIs probabilistically based on all the observed data (including the multimapped reads) in each cell.

With regard to capturing multi-gene mapping reads (impacting ~2000 and ~5000 genes in mice and humans, respectively), our manuscript outlines a conclusive solution for this problem by fixing overlap causing transcripts. Prior solutions (e.g. StarSolo's maximum likelihood estimation) assigns multigene mapping reads between genes where in most cases, no read sharing is required and can unambiguously be assigned. There are also other frameworks implemented by StarSolo for capturing multigene mapping reads (Rescue, Propuniqu, Uniform) which come with rigid assumptions that do not hold true for most overlapping gene cases.

We will compare StarSolo's implementation of multimapping read mapping resolution strategies (EM, Rescue and Uniform) to our solution and provide that as a supplemental figure.

Therefore, it would be useful if the authors could compare the results from the optimized reference transcriptome with those from the existing pipelines by setting the flag of including intronic reads and multimapping reads. For instance,in STARsolo, the user should specify --soloFeatures GeneFull and --soloMultiMappers EM. Avery direct and

meaningful comparison would be to show e.g. the effect of just removing the overlapping genes versus resolving the multimapping reads using theEM algorithm — is removing the overlappers necessary? Does it perform betterthan STARsolo's probabilistic resolution?

We agree that such comparison is useful to assess how each optimization step contributes to gene recovery. We will provide this analysis in Extended Figure 2A.

Figure 5 is informative, clean, and quite striking. However, it is hard to understand the result without the code or a detailed description of the analysis in the manuscript.

Will specify parameters in Methods and provide code with input data files on Github.

To be specific, it is not clear if the new clusters (not cell types, the clusters returned by the Seurat `FindClusters()` function) can be detected by increasing the resolution in the exonic-reference-based analysis; it is not clear if the cells in each cluster after expanding the reference are the same as before; it is not clear if the newly discovered genes can be detected as the highly variable genes by the Seurat `FindMarkers()` function to be used for identifying cell types in the standard Seurat pipeline (https://satijalab.org/seurat/articles/pbmc3k_tutorial.html).

For both mouse brain and human PBMC datasets, increasing the clustering resolution with the exonic reference does not recover several known cell types revealed by the optimized reference as the required differentially expressed genes are not included in the list of differentially expressed genes. We will include a supplemental figure showing the location of newly detected cell class members in the exonic reference analysis to illustrate this point.

It is unclear whether the newly discovered genes are the only markers of the newly discovered cell types. I would like to suggest the authors upload an R or Python notebook of the analysis of generating Figure 5 and have an explanation of the process of cell type assignment in the Methods section.

We will include an R Markdown document on Github for generating Figure 5 as well as the input data and expand the explanation in the Methods section.

A follow-up suggestion is that because the marker genes included in Figure 5 are very important, it will be useful to show where the reads of those genes map and by which mechanism (intronic reads, unannotated 3' UTR, or overlap genes) those marker genes failed to be recovered by the exonic reference.

Will include in supplemental figure.

The optimized reference transcriptome has the potential to become a key step in standard single-cell analysis. However, according to the codebase in the provided GitHub repository, the current implementation can't be directly utilized by the community in a straightforward and generalizable manner. The usability of the method itself is likely to be quite important to its adoption. I strongly suggest the authors develop a properly-documented and easily-installable self-contained package for the construction of the optimized reference transcriptome. Ideally, this package could be submitted to Bioconductor or CRAN, or if implemented in python, could be distributed via pip or bioconda. At the very least, the code should be made into an R package, with all required dependencies, installable from its GitHub repository via `devtools` and proper documentation and vignette should be provided. Additionally, some tutorials and example runs demonstrate using the package to generate and quantify against the optimized reference transcriptome would be helpful to potential users, though such additional resources aren't strictly necessary if a full package with documentation is available.

This is an excellent suggestion. We will generate an R package that can be deployed with devtools from Github repository and generate R vignettes to make this approach easily accessible for the broader community. Considering that mouse and human genome annotations are continuously updated and occasionally new genome builds are released, this will likely be highly useful. We will also keep the commented source code on Github to facilitate customization.

G. References: appropriate credit to previous work?

The authors gave credit to almost all previous work. One small suggestion is that the author could mention that many existing methods that have the ability to process intronic reads, like CellRanger, kallisto|bustools, STARsolo, zUMIs, alevin,and alevin-fry.

Will add.

H. Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction, and conclusions

Overall, the manuscript is well written and easy to follow. The authors clearly expressed the rationale of the work, the details of the implementation, and the contribution and significance of the proposed optimized reference transcriptome. I believe the authors have proposed some straightforward and concrete improvements to the reference transcriptome curation process that has the potential to positively impact a broad variety of single-cell sequencing analysis studies.

Thank you.

Decision Letter, third revision:

30th Oct 2022

Dear Dr Oka,

Thank you for your letter asking us to reconsider our decision on your Article, "Enhanced recovery of single-cell RNA-sequencing reads for missing gene expression data". After careful consideration we have decided that we are willing to consider a revised version of your manuscript that fully addresses all the concerns raised by our reviewers.

When revising your paper:

* include a point-by-point response to our referees and to any editorial suggestions

* please underline/highlight any additions to the text or areas with other significant changes to facilitate review of the revised manuscript

* address the points listed described below to conform to our open science requirements

* ensure it complies with our general format requirements as set out in our guide to authors at www.nature.com/naturemethods

* resubmit all the necessary files electronically by using the link below to access your home page

[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.

We hope to receive your revised paper within eight weeks. We are very aware of the difficulties caused by the COVID-19 pandemic to the community. If you cannot send it within this time, please let us know. In this event, we will still be happy to reconsider your paper at a later date so long as nothing similar has been accepted for publication at Nature Methods or published elsewhere.

OPEN SCIENCE REQUIREMENTS

REPORTING SUMMARY AND EDITORIAL POLICY CHECKLISTS
When revising your manuscript, please submit reporting summary and editorial policy checklists.

Reporting summary: https://www.nature.com/documents/nr-reporting-summary.zip
Editorial policy checklist: https://www.nature.com/documents/nr-editorial-policy-checklist.zip

If your paper includes custom software, we also ask you to complete a supplemental reporting summary.

Software supplement: https://www.nature.com/documents/nr-software-policy.pdf

Please submit these with your revised manuscript. They will be available to reviewers to aid in their evaluation if the paper is re-reviewed. If you have any questions about the checklist, please see http://www.nature.com/authors/policies/availability.html or contact me.

Please note that these forms are dynamic 'smart pdfs' and must therefore be downloaded and completed in Adobe Reader. We will then flatten them for ease of use by the reviewers. If you would like to reference the guidance text as you complete the template, please access these flattened versions at http://www.nature.com/authors/policies/availability.html.

DATA AVAILABILITY
Please include a "Data availability" subsection in the Online Methods. This section should inform readers about the availability of the data used to support the conclusions of your study, including accession codes to public repositories, references to source data that may be published alongside the paper, unique identifiers such as URLs to data repository entries, or data set DOIs, and any other statement about data availability. At a minimum, you should include the following statement: "The data that support the findings of this study are available from the corresponding author upon request", describing which data is available upon request and mentioning any restrictions on availability. If DOIs are provided, please include these in the Reference list (authors, title, publisher (repository name), identifier, year). For more guidance on how to write this section please see: http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf

CODE AVAILABILITY
Please include a "Code Availability" subsection in the Online Methods which details how your custom code is made available. Only in rare cases (where code is not central to the main conclusions of the

paper) is the statement “available upon request” allowed (and reasons should be specified).

We request that you deposit code in a DOI-minting repository such as Zenodo, Gigantum or Code Ocean and cite the DOI in the Reference list. We also request that you use code versioning and provide a license.

For more information on our code sharing policy and requirements, please see: https://www.nature.com/nature-research/editorial-policies/reporting-standards#availability-of-computer-code

MATERIALS AVAILABILITY
As a condition of publication in Nature Methods, authors are required to make unique materials promptly available to others without undue qualifications.

Authors reporting new chemical compounds must provide chemical structure, synthesis and characterization details. Authors reporting mutant strains and cell lines are strongly encouraged to use established public repositories.

More details about our materials availability policy can be found at https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards#availability-of-materials

ORCID
Nature Methods 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. This applies to primary research papers only. 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 hearing from you soon. Thank you very much for submitting your paper to Nature Methods.

Sincerely,

Lin Tang, PhD

nature research

Senior Editor
Nature Methods

Author Rebuttal, third revision:

Response to reviewers

We thank the reviewers for the constructive and thoughtful comments and suggestions. Below, we have outlined the changes made in the revised version of the manuscript and addressed the reviewers' individual remarks. We are grateful for the possibility of revising the manuscript and hope the reviewers will now find it suitable for publication in Nature Methods.

Reviewer 1:

The manuscript by Pool et al. makes the case for refining the gene annotations used in scRNA-seq analysis to capture more signal in data generated using 3'protocols. Currently, various read processing methods allow the inclusion of intron reads by default or as an option in the gene counting step (e.g. zUMIs https://pubmed.ncbi.nlm.nih.gov/29846586/, alevin-fry (spliced + intron reference,https://www.nature.com/articles/s41592-022-01408-3) - please consider citing these works), CellRanger v7 https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/release-notes).

In the Discussion section we now include references to the above as well as other aligners (Kallisto and STARsolo) that enable incorporation of intronic reads and we refer to work benchmarking their efficiency (Du et al. 2020). We also now show how our hybrid pre-mRNA incorporation approach as well as the overall optimized reference based solution outperforms the existing intronic read incorporation approaches (Supplemental Data Figure 3, Figure 5a, b).

However, reads lost to 3′ UTR annotation issues or overlapping features has not been systematically explored to my knowledge so bringing this altogether to look at their relative contribution to signal is valuable.

General questions: How tissue specific are these results i.e. does the relative contribution of signal from the 3 sources vary depending upon whether you are looking at brain, blood vs other tissues? This study uses mouse brain tissue and a PBMC sample, which is a good start, but given the large amount of public scRNA-seq data available, it should be relatively easy to explore this in other sample / cell types (perhaps an atlas dataset?) to assess how generalisable these findings are.

This is an important suggestion. We evaluated the tissue specific effects on gene proximal 3′ intergenic read prevalence in a panel of mouse and human tissues (five total per species using publicly available scRNA-seq data mostly generated by the Tabula Muris and Tabula Sapiens consortia) and found that less than a quarter of genes with large 3′ proximal intergenic read mapping are shared across all tissues. These results suggest that substantial gains could still be made by further improving 3′ gene end annotations by analyzing read mappings across diverse tissues. The results of these analyses are now presented in new Figure 2d and e.

Does the optimized annotation have to be redefined for every experiment /different scRNA-seq protocols, or is it generalisable?

In general most analyses would not require generation of a new reference for mouse and human data as gene overlap resolution and the hybrid intronic read capture strategy do not require any further customization. A large portion of the 3′ gene end correction is also broadly beneficial but may benefit from further gene end modifications building on top of existing corrections which we

have already performed. The latest version of the optimized references clearly outperform both exonic and exon+intron references with respect to number of genes detected and reads registered across diverse tissues (see new Figure 5a, b, Supplemental data figure 3). These results argue that the optimized references are beneficial as is and can be used to gain additional biological insight across broad tissue contexts.

Depending on the user and specific tissue, some modification may be desirable. To avoid the need to repeat the amount of hands on curation, we have provided the input files for generating the optimized human and mouse references that other users may easily modify for the benefit of their analyses (available at www.thepoollab.org/resources and https://github.com/PoolLab/generecovery). For example, although a large part of fixing the 3' gene ends has already been done by us, different tissues are likely to benefit from follow-up improvement in 3' gene boundaries as is evident from our analysis in Fig 2e, d.

How feasible is the case-by-case scrutiny for recovering intergenic reads from 3' unannotated exons and how long does this step take? If these signals are highly dataset specific (i.e. only relevant for brain mouse samples and PBMC human samples) and require significant optimization, this may limit the usefulness of the online resources provided by the authors.

Case by case scrutiny and quantification is feasible but takes quite a bit of time (~15 minutes per gene in the candidate list for an experienced user). We have already done that for thousands of genes via manual curation and scrutiny of analyzed datasets. With specific datasets, users can already build on our work and only curate genes that we did not find in our prioritized gene list for our data but which come up in their gene list for genes with large 3' intergenic read loadings. Also, our fixed 3' gene list includes best estimates of 3' gene ends from Refseq and Ensembl annotations.

In order to estimate, how much could individual users benefit from additional gene end curation, we looked at how many genes show high levels of tissue specific 3' proximal intergenic read loading and found that this varies between 15-20% of top 1000 genes. Thus, customized gene curation could be useful. However, we would like to stress that such custom curation can be easily built upon our database.

Of the interventions suggested, it appears that intron inclusion boosts signal the most (Fig 1e and g), although many new genes are added through both intron or 3'intergenic mapping (Fig 1f). I wondered about the relative contribution of these new genes in downstream analysis though. Are many of the newly detected genes highly variable and drivers of the new clustering results, or are they mostly lowly expressed / uninformative and likely to be filtered out from downstream analysis as part of quality control? Or perhaps the two analyses (exonic reference vs optimizedreference) are based on signal from mostly the same genes, just with enhanced signal that has improved cell type specificity in the reference optimized results? It would be good to do some analysis to tease this apart to provide the reader with further insights on how many of the 'missed' features are likely to turn up as genes of interest (hyper-variable / marker genes). A number of examples are provided in Fig 5 and Extended Fig 3, but it wasn't clear if these were it, or if there are any more such examples.

We looked at the source of genes that comprise this newly contributing gene set and found that about 30-40% of these genes are directly the result of optimization strategies employed in our paper manuscript (i.e. genes with fixed 3' gene ends, overlapping genes or genes showing significantly more reads due to the hybrid pre-mRNA detection strategy as compared to regular intronic reference). We added these data in Supplemental figure 3. We also analyzed that directly in both mouse MnPO and human PBMC datasets (see new Extended Data Figure 3A) and found that the use of optimized reference results in about half (46-54%) of the variable genes to be replaced by new ones. This

suggests that the increased quality in cell-type and -state resolution is the result of incorporating new biological signal from genes that otherwise would not impact the clustering. In sum, the recovered genes are not necessarily lowly expressed genes, but instead impactful ones with respect to extracting biological insight.

What are the signal / biotype characteristics of the unique genes in Extended Fig 2i.e. are they lowly expressed features, or do they span the intensity range? Are particular gene biotypes over-represented in this category?

We tested for gene biotypes that are uniquely detected by different intronic read icorporating methods using Gene Ontology term enrichment analysis with ClusterProfiler package in R but did not find any enriched terms. This likely reflects lack of biological function bias displayed by different methods. We also looked at the expression range of genes selectively detected by specific intronic read incorporating methods and found that while the majority are low expressing genes, many such genes (>100) span the expression range from intermediate to high (new Extended Data Figures 2d, e).

Specific comments:

- *Consider combining Figures 1 and 2?*
- *Consider providing a CHM13 version of the human annotation?*

As we have revised Figure 2 to include additional panels, combining Figures 1 and 2 has now become unpractical due to space constraints. We would therefore prefer to keep them separate.

We considered generating a CHM13 version, however this would end up providing 2 competing references for human data that can be confusing for users. We decided to keep the versions based on reference genomes made

available by 10x genomics (most widely used and enable direct comparison). We will generate updated versions once 10x genomics transitions to newest genome builds.

I'm not sure that the claim in the sentence that begins on line 292 with 'These approaches…' is correct. It was my understanding that imputation methods are often applied to 3' scRNA-seq data.

Thank you for pointing this out. We have now adjusted this section to more accurately reflect the content and limitations of imputation.

Typos on line 298, 424, 450, 477

Thank you, these are now fixed.

Add more detail to Methods on line 674 (e.g. what options were used for filtering poor quality cells / genes with low abundance / normalization / clustering in Seurat? Given the results in Figure 5, the methods here seem incomplete).

We have now extended the Methods section to reflect the content of the analysis (under subheading "Analysis of MnPO and PBMC scRNA-seq data for benchmarking optimized reference performance") as well as uploaded the code to our Github repository (https://github.com/PoolLab/generecovery/tree/main/Data_analysis/Figure5).

Reviewer #2:

A. Summary of the key results

This paper introduced a workflow for the construction of an optimized transcriptomic reference for improving the profiling resolution (specifically, increasing the sensitivity of detected genes) of single-cell RNA-sequencing data analysis. In the proposed workflow, to build the optimized reference transcriptome, the following steps are taken:
1. The exonic reference is replaced by a pre-mRNA reference
2. Overlapping genes are removed if satisfying certain criteria
3. The 3' UTR of qualified genes are extended

The key results are:
Quantitative evaluation of the three main mechanisms that cause loss of reads in the traditional exonic-reference-based analysis. Demonstration of a general strategy to build an optimized reference transcriptome. Description and evaluation of the benefits of replacing the traditional exonic reference with the optimized reference transcriptome in single-cell data analysis. Specifically, the demonstration that thousands of genes absent in the exonic-reference-based analysis can be recovered, and new cell types are identified by those newly recovered genes in both a human and a mouse dataset.

B. Originality and significance: if not novel, please include reference

The key approaches introduced in this manuscript extend, expand upon and refine a previous paper published by the same first author, "Pool, A.-H. et al. The cellularbasis of distinct thirst modalities. Nature (2020). " In the previous paper, the expanded transcriptome is built by including unspliced transcripts, removing 294 pseudogenes with little to no exonic read mapping from the genome, and extending 18 genes' untranslated regions. In the current manuscript, the construction process of the expanded transcriptome is expanded and formalized into a much more general computational workflow. The authors implemented the workflow in R so that users can construct the optimized transcriptome in a partially automated way (though with the help of some manual inspection). The

authors also provided an optimized transcriptome reference for mouse and human. Part of the proposed workflow, specifically, the inclusion of intronic reads in the reference, has been adopted by some existing quantification tools. For example, Cell Ranger 7 includes the intronic reads in the quantification process by default (see https://support.10xgenomics.com/docs/intron-mode-rec?src=social&lss=linkedin&cnm=soc-li-ra_g-program&cid=7011P000000y072). Starsolo (Kaminow et al., bioRxiv, 2021) has the option `--soloFeatures GeneFull` to include the intronic reads in its procedure, as described in section 5.1.3 in starsolo preprint. The splici (spliced + intronic) transcriptomic reference used in alevin-fry (He et al.,Nature Methods, 2022) contains the introns of genes, as demonstrated in Supplementary section S2 in its manuscript. Kallisto bustools (Melsted et al., Nat Biotech, 2021) can take into account intronic reads if making corresponding changes and passing `--workflow lamanno` in its pipeline.

An important takeaway of the paper is that the precise strategy for including intronic reads has a significant impact on which genes are detected and how efficiently (Extended data Figure 2). Our hybrid intronic read capture solution that entails combining the regular intronic read inclusion with the addition of gene length exons captures close to 400 extra genes that otherwise go undetected (Extended Data Figure 2 a, b) and allows us to detect 50% or more reads from thousands of genes, depending on the dataset. We have quantified the improvements stemming from this approach in Extended Data Figures 2, 3 and compared the optimized reference to regular intronic reference in Fig. 5 a, b.

On top of these existing methods, the significance of this manuscript comes from the fact that this optimized reference transcriptome further reduced the loss of reads, compared with the existing tools, by removing low-quality overlapping genes and extending the 3' UTR of genes that are considered as having unannotated 3' UTR region. These two strategies have rarely been explored and

have the potential to be adopted as a standard step for single-cell data processing pipeline. Moreover, the author provided the R scripts for building the optimized reference transcriptome for any given genome.

C. Data & methodology: validity of the approach, quality of data, quality of presentation

The Methods section was well written. It is easy to follow the text and the corresponding figures to understand the process of building an optimized reference transcriptome. The only exception is that the algorithm in the "Resolving gene overlap derived read loss" section is a bit hard to follow. It would be great if a figure plus a formal algorithmic style description (e.g. more in the style of pseudocode) could be added to help demonstrate the process.

We rewrote this section in pseudocode style and simplified the description of steps to make it easier to follow. We did attempt to summarize the process in a figure but due to the multistep nature of this procedure, the pseudocode is more straight-forward to relay the necessary steps.

Additionally, I have the following concerns. The GitHub repository provided in the manuscript, https://github.com/PoolLab/Generecovery, doesn't contain all the code of the steps described in the manuscript. For example, in the file 1_Annotation pre-processing.R, step A: resolve gene overlap derived read loss only contains the code for step A1:Generate a rank-ordered gene list of same-strand overlapping genes, but not A2: Prioritize gene list by classifying overlapping into recommended action categories.

R implementation of the section A2 has now been added to Github.

Moreover, In the provided R script, bash commands also exist, and some strings (gene names and paths) are hard-coded in the script, making the workflow hard

to apply by the user and near impossible to reproduce precisely. At the very least, any dependent or necessary files should be included in the repository and referenced by an appropriate relative path so that the scripts can easily be adopted by a user. Please see section F below for a more detailed suggestion.

We have updated the R scripts in the Github repository. Specifically, we have converted all python and bash portions of the script into R and the remaining bash sections are invoked from R. Also, we have got rid of all the hard-coded elements in the script and replaced them with flexible options to introduce required input data from files. We have also included all the necessary input files for generating the references on the Github repository (previously they were available from at the www.thepoollab.org/resources, which may have been confusing).

The mouse and human datasets included in the paper clearly show that the three problems of the current reference are outstanding and need to be addressed. Figure 5 shows that the optimized reference transcriptome is able to recover thousands of genes and therefore improve the cell type identification result. One concern is that the code used for generating figure 5 is missing in the GitHub repo. Meanwhile, in the Methods section, the process of clustering and labeling the cells is unclear. For example, the parameters used in the clustering analysis, such as the number of variable features and PCs used for initial dimensionality reduction, the number of nearest neighbors for computing k-NN graph, and the resolution for the clustering algorithm (if following the standard Seurat pipeline,https://satijalab.org/seurat/articles/pbmc3k_tutorial.html), are not specified.

We have extended the Methods section to reflect the content of the analysis (under subheading "Analysis of MnPO and PBMC scRNA-seq data for benchmarking optimized reference performance") as well as uploaded the code

to the Github repository (https://github.com/PoolLab/generecovery/tree/main/Data_analysis/Figure5). The analysis parameters for the comparative analysis were identical (same number of variable genes, PC-s, clustering resolution for exonic and optimized reference datasets).

D. Appropriate use of statistics and treatment of uncertainties
These were used appropriately.

E. Conclusions: robustness, validity, reliability

The section is clear and easy to follow. It demonstrates the problem that this work wants to address and the contributions this work makes.

F. Suggested improvements: experiments, data for possible revision

It is not clear from the paper how much of the gains of replacing the exonic transcriptomic reference with the optimized reference transcriptome in the downstream analyses are from the inclusion of the introns, which has been adopted by several other methods, and how much additional is gained from removing overlapping genes and extending 3' UTR of genes. If possible, a follow-up experiment can be added to show the importance of the newly recovered genes by addressing each of the three problems for the potential downstream analyses.

We evaluated the relative contribution of all three improvement strategies (hybrid pre-mRNA based intronic read capture, 3' gene extensions and elimination of gene overlaps) in a new Extended Data Figure 3. We see that the use of optimized references results in about 50% of variable genes being replaced by new genes (Extended Data Figure 3A) with 30-40% of the new genes being the result of hybrid pre-mRNA derived enhanced intronic read capture,

resolving of overlapped reads or 3' gene extension. Predictably, the proportion of each contribution various by dataset.

Furthermore, we evaluated the optimized reference mapping and individual improvement strategy derived increased read recovery and new gene detection as well as biased read capture in comparison to exonic and exon+intron read integrating references in Extended Data Figure 3b-g on a wide variety of mouse and human scRNA-seq datasets (n=5 per species, mostly Tabula Muris and Tabula Sapiens scRNA-seq datasets).

For the genes recovered by addressing each of the three problems, how many of them are protein-coding genes?

This is highly dependent on the dataset being analyzed but more than 75% are protein coding when we compare genes recovered with optimized reference with intronic or exonic reference mapped data.

How many of them participate in important pathways of the studied samples? How many of them are the cell type markers of the studied samples? Are those genes significantly differentially expressed in one of the newly discovered clusters?

Many of the genes improved by the different optimization strategies we implement play important roles in the profiled tissue. In Figure 5g we highlight important PBMC marker genes with colored triangles that show significantly improved detection as a result of optimization (blue triangles label genes with significantly improved detection thanks to 3' gene end detection; red triangles indicate genes gaining mostly from additional intronic read incorporation). Furthermore, Figure 5d and Extended data figure 4 display improvements and marker genes for the mouse MnPO analysis including B4galnt2 which we detect only thanks to hybrid pre-mRNA capture strategy but not via regular intronic capture nor exonic reference mapping approaches. To visualize the proportion

of optimized reference derived new genes determining cell clustering, we have now included Extended Data Figure 3A for mouse MnPO and human PBMC datasets. We also plotted additional cell-class markers stemming from resolving gene overlaps and 3′ gene extensions in Extended Data Figure 4.

Are all of these additional changes necessary to discover the new cell types / clusters, or do just introns or just extended 3′ UTRs suffice? The purpose of asking this follow-up experiment is to show the importance of addressing each of the three problems for the proposed improvement in the paper.

The majority of improvement comes from including intronic reads, followed by resolving gene overlap derived read loss and then integrating intergenic reads in unannotated 3′ UTRs. We have now quantified the gains from each of the improvement strategies in a new Figure (Extended Data Figure 3) with respect to extra genes and reads detected. Furthermore, we have quantified the gains from using the hybrid-premRNA strategy for capturing intronic reads in updated and new figure panels (Extended Data Figure 2). Again the relative contribution of each of these strategies depends on the profiling resolution of cells as well as the specific biological sample. For the MnPO and PBMC datasets, we are now showing the relative proportion of new variable genes stemming from the three optimization strategies for our optimized references (insets in Extended Data Figure 3A) which shows that roughly 30-40% of new variable genes in these analyses derive from issues addressed by our optimization (hybrid strategy for capturing intronic reads, gene overlaps and 3′ intergenic reads). As variable genes determine cellular clustering, these likely contribute proportionally to cell-type specific markers as the rest of variable genes.

Another suggestion is that the existing pipelines like STARsolo have the ability to include intronic reads in the final cellular gene expression data by specifying flags and the ability to solve the multimapping reads by an EM framework.

We have compared STARsolo's GeneFull method to Cell Ranger –include-introns mode and our hybrid-pre-mRNA reference in Extended Data Figure 2a-c. The EM framework ends up assigning probabilities for transcripts to stem from certain genes if the map to multiple genes. This is not directly comparable to unambiguously assigned reads in our solution where we eliminate the overlap causing transcripts enabling us to assign otherwise discarded reads.

The inclusion of the intronic reads corresponds to the inclusion of the unspliced transcripts in the current manuscript. Likewise, the EM workflow is designed to solve the loss of reads caused by overlapping genes because, instead of discarding the multimapping reads, the EM algorithm will try to appropriately assign UMIs probabilistically based on all the observed data (including the multimapped reads) in each cell.

With regard to capturing multi-gene mapping reads (impacting ~2000 and ~5000 genes in mice and humans, respectively), our manuscript outlines a conclusive solution for this problem by fixing overlap causing transcripts and non-coding genes. Prior solutions (e.g. StarSolo's EM workflow), while very useful, assign multigene mapping reads between genes where in most cases, no read sharing is required and can unambiguously be assigned by deleting the offending read-through or premature transcripts. There are also other frameworks implemented by StarSolo for capturing multigene mapping reads (Rescue, Propunique, Uniform) which come with assumptions that often do not reflect biological reality. These solutions also end up diluting the biological signal between many genes (e.g. >10 genes for Ugt1, Pcdha and Pcdhgb genes in the mouse genomes) where no dilution is required. Therefore the solution we put forward is a better way of capturing biological reality and signal. We have added a short sectioning addressing these issues in the Discussion section.

Therefore, it would be useful if the authors could compare the results from the optimized reference transcriptome with those from the existing pipelines by

setting the flag of including intronic reads and multimapping reads. For instance,in STARsolo, the user should specify --soloFeatures GeneFull and --soloMultiMappers EM. Avery direct and meaningful comparison would be to show e.g. the effect of just removing the overlapping genes versus resolving the multimapping reads using theEM algorithm — is removing the overlappers necessary? Does it perform betterthan STARsolo's probabilistic resolution?

We have compared STARsolo's GeneFull method to Cell Ranger –include-introns mode and our hybrid-pre-mRNA reference in Extended Data Figure 2a-c.

Figure 5 is informative, clean, and quite striking. However, it is hard to understand the result without the code or a detailed description of the analysis in the manuscript.

We have now extended the Methods section to reflect the content of the analysis (under subheading "Analysis of MnPO and PBMC scRNA-seq data for benchmarking optimized reference performance") as well as uploaded the code to our Github repository (https://github.com/PoolLab/generecovery/tree/main/Data_analysis/Figure5).

To be specific, it is not clear if the new clusters (not cell types, the clusters returned by the Seurat `FindClusters()` function) can be detected by increasing the resolution in the exonic-reference-based analysis; it is not clear if the cells in each cluster after expanding the reference are the same as before; it is not clear if the newly discovered genes can be detected as the highly variable genes by the Seurat `FindMarkers()` function to be used for identifying cell types in the standard Seurat pipeline (https://satijalab.org/seurat/articles/pbmc3k_tutorial.html).

There are two types of newly uncovered clusters that we observe as a result of using the optimized references:

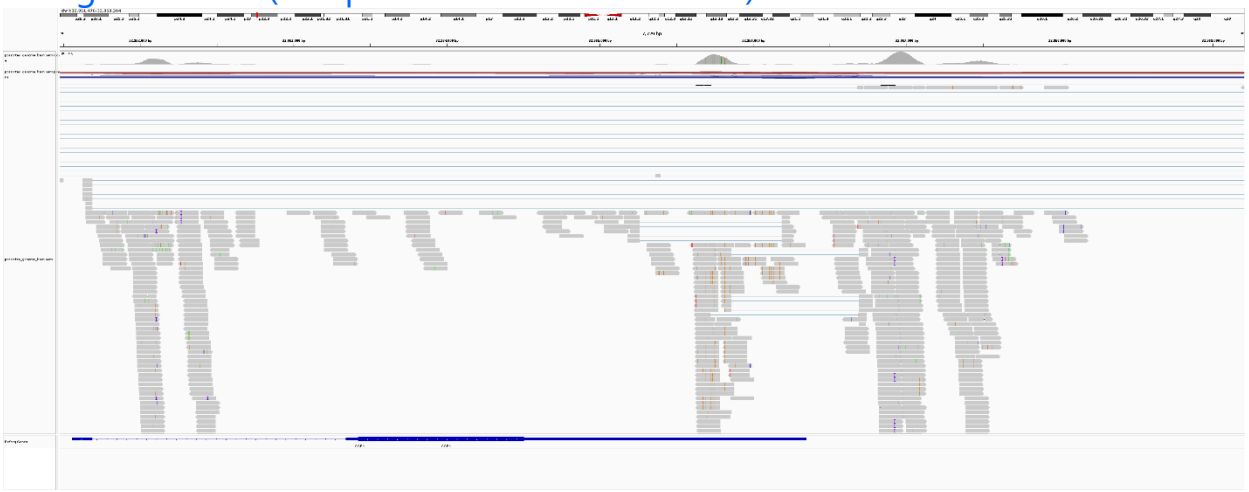
- Cell-types and clusters that do not get segregated out with the exonic reference regardless of input variable gene numbers and clustering resolution as it lacks required gene expression data to reveal the cell-type. Separation of the third Ptger3+ Etv1+ Glut3 neuron type in the MnPO (Extended Data Figure 4b) with the optimized reference is an example of that which does not separate out with exonic reference as it fails to detect Ptger3 expression and other genes defining this cluster. We know that it is a biologically relevant neuron-type as there is a homologous neuron type in the related thirst behavior mediating brain nuclei OVLT and SFO that express identical markers (Etv1+, Ptger3+) and function as mineral and salt appetite mediating neurons (Pool et al. Nature. 2020, Zhang, Pool et al. submitted).
- Cell-types that would eventually segregate out but that lack information on canonical markers and generate a large number of noise/nonsense clusters at the required clustering resolution. T-regs in the human PBMC dataset (Figure 5f) are a case in point that with exonic reference lack their characteristic marker expression (IL2RA).
- We added a short discussion of these issues in the discussion section.
- We agree, that it is somewhat challenging to keep track of which cell types correspond to which in the exonic and optimized reference mapped data in Figure 5 and Extended Figure 4. We have aligned them such that new cell-types appear from the same "source" cluster in the optimized reference and aligned the cluster names accordingly (Extended Data Figure 4a, b).

It is unclear whether the newly discovered genes are the only markers of the newly discovered cell types. I would like to suggest the authors upload an R or Python notebook of the analysis of generating Figure 5 and have an explanation of the process of cell type assignment in the Methods section.

No, there are plenty of other cell-type specific markers that get revealed by using the enhanced reference. We have now included the input data and the used code on Github for easy reproduction (https://github.com/PoolLab/generecovery/tree/main/Data_analysis/Figure5). We have also clearly outlined the analysis conducted in the Methods section under the subheading "Analysis of MnPO and PBMC scRNA-seq data for benchmarking optimized reference performance".

A follow-up suggestion is that because the marker genes included in Figure 5 are very important, it will be useful to show where the reads of those genes map and by which mechanism (intronic reads, unannotated 3' UTR, or overlap genes) those marker genes failed to be recovered by the exonic reference.

We have included this analysis in the Extended Data Figure 4 for the mouse MnPO data (plotted read mapping in Gdf11 – 3' intergenic read loading data, Ptger3 – labeling thirst and thermoregulation relevant cell populations from intronically mapped read data and Kisspeptin gene recovered from gene overlap resolution). The same applies for human PBMC data – e.g. Th2 marker chemokine receptor 4 (CCR4) that results in 3x increased reads thanks to 3' intergenic reads (IGV plot below for illustration).



The optimized reference transcriptome has the potential to become a key step in standard single-cell analysis. However, according to the codebase in the provided GitHub repository, the current implementation can't be directly utilized by the community in a straightforward and generalizable manner. The usability of the method itself is likely to be quite important to its adoption. I strongly suggest the authors develop a properly-documented and easily-installable self-contained package for the construction of the optimized reference transcriptome. Ideally, this package could be submitted to Bioconductor or CRAN, or if implemented in python, could be distributed via pip or bioconda. At the very least, the code should be made into an R package, with all required dependencies, installable from its GitHub repository via `devtools` and proper documentation and vignette should be provided. Additionally, some tutorials and example runs demonstrate using the package to generate and quantify against the optimized reference transcriptome would be helpful to potential users, though such additional resources aren't strictly necessary if a full package with documentation is available.

We thank the reviewer for this important suggestion. We generated an R package *ReferenceEnhancer* that is installable from Github repository with devtools (available from https://github.com/PoolLab/ReferenceEnhancer with links provided at www.thepoollab.org/resources). The package streamlines the gene prioritization for gene overlap resolution, 3' gene extension as well as optimized transcriptome annotation generation. We have also provided documentation and vignettes with sample data on Github to facilitate easy pickup for the method.
Considering that genome annotations are continuously updated and occasionally new genome builds are released, this will likely be a highly useful resource to enable the community to extract most insight from their single-cell data. We will also keep the commented source code on Github to facilitate customization.

G. References: appropriate credit to previous work?

The authors gave credit to almost all previous work. One small suggestion is that the author could mention that many existing methods that have the ability to process intronic reads, like CellRanger, kallisto|bustools, STARsolo, zUMIs, alevin,and alevin-fry.

References have been added.

H. Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction, and conclusions

Overall, the manuscript is well written and easy to follow. The authors clearly expressed the rationale of the work, the details of the implementation, and the contribution and significance of the proposed optimized reference transcriptome. I believe the authors have proposed some straightforward and concrete improvements to the reference transcriptome curation process that has the potential to positively impact a broad variety of single-cell sequencing analysis studies.

Thank you.

Decision Letter, fourth revision:

Our ref: NMETH-A46722D

17th Apr 2023

Dear Dr. Oka,

Thank you for submitting your revised manuscript "Enhanced recovery of single-cell RNA-sequencing reads for missing gene expression data" (NMETH-A46722D). 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 Methods, pending minor revisions to satisfy the referees' final requests and to comply with our editorial and formatting guidelines.

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

TRANSPARENT PEER REVIEW

Nature Methods offers a transparent peer review option for new original research manuscripts submitted from 17th February 2021. We encourage increased transparency in peer review by publishing the reviewer comments, author rebuttal letters and editorial decision letters if the authors agree. Such peer review material is made available as a supplementary peer review file. **Please state in the cover letter 'I wish to participate in transparent peer review' if you want to opt in, or 'I do not wish to participate in transparent peer review' if you don't.** Failure to state your preference will result in delays in accepting your manuscript for publication.

Please note: we allow redactions to authors' rebuttal and reviewer comments in the interest of confidentiality. If you are concerned about the release of confidential data, please let us know specifically what information you would like to have removed. Please note that we cannot incorporate redactions for any other reasons. Reviewer names will be published in the peer review files if the reviewer signed the comments to authors, or if reviewers explicitly agree to release their name. For more information, please refer to our FAQ page.

ORCID
IMPORTANT: Non-corresponding authors do not have to link their ORCIDs but are encouraged to do so. Please note that it will not be possible to add/modify ORCIDs at proof. Thus, please let your co-authors know that if they wish to have their ORCID added to the paper they must follow the procedure described in the following link prior to acceptance:
https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research

Thank you again for your interest in Nature Methods. Please do not hesitate to contact me if you have any questions. We will be in touch again soon.

Sincerely,

Lin Tang, PhD
Senior Editor
Nature Methods

Reviewer #1 (Remarks to the Author):

The have authors have adequately addressed each of my questions in their revised manuscript.

Reviewer #2 (Remarks to the Author):

We thank the authors for addressing all of the comments and suggestions and answering all of the questions we raised in the review.

Overall, from the rebuttals and updated manuscript, it is clear that all three improvements of the optimized reference over traditional spliced transcriptome reference contribute to the increased gene detection rate and the newly discovered cell markers.

Questions and suggestions:

1. In the response to reviewers, the authors said that they ran STARsolo with the EM algorithm, but in the method section, we did not see --soloMultiMappers EM anywhere. We suggest the authors make explicit mention of how STARsolo was run with the EM, and update the command listed in the methods section to include this option, as the result was discussed thoroughly in the manuscript.

2. It is unclear specifically what the "intronic reference" refers to in Fig 5(b) title. Please clarify it in the caption.

3. In Ext Fig 3(a), we suggest changing the color of the bar that represents "regular intronic" from black to another color, as black seems to be assigned multiple meanings in the same plot.

4. In Ext Fig 2 and line 178, It is unclear why adding a transcript length exon (hybrid pre-mRNA reference) for each transcript improves the result of –include-introns. It will be useful if the authors can briefly explain this somewhere in the text, as the difference of this approach from the standard "geneFull" –-soloFeatures flag is not clear.

5. In Ext Fig 2 (a), the leftmost label on the x-axis, "Regular pre-mRNA reference", should be changed to "Cell Ranger pre-mRNA reference".

6. In Figure 1(f), it is obvious that the newly detected genes are found by adding introns and extending 3' UTRs overlap. Still, the difference is difficult to see without the exact number of yellow and pink bars. Therefore, we suggest the authors add the number represented by these two bars, just like the other two in the same plot.

7. In line 24: It is unclear if each of these issues impacts the detection of thousands of genes from Extended Fig 3.

8. It is unclear how the terminology "readthrough transcript" is used in the manuscript, and it is not exactly the same as what is introduced in an Ensembl blog (https://www.ensembl.info/2019/02/11/annotating-readthrough-transcription-in-ensembl/). We suggest the authors clarify this in the text.

9. In line 304, reference #41 only compared STARsolo and Kallisto but not other methods. The conclusion from that paper was STARsolo outperforms Kallisto w.r.t. detecting genes, memory usage, and # of mapped reads. It is unclear how that conclusion is related to the point discussed in the current manuscript.

10. In the "Recovering intronic reads" section (Line 678), the authors introduced three ways to build an intronic reference. The traditional approach, the approach adopted by STARsolo and Cell Ranger, and the new hybrid approach. It is unclear which one was compared in Figure 5 (a) and (b) and Line 306 in the Discussion section and was called "regular exonic and intronic read incorporating reference". We suggest the authors clarify it somewhere in the text or in the caption of Figure 5.

11. In the section Analysis of MnPO and PBMC scRNA-seq data for benchmarking optimized reference performance, the authors clearly explained the procedure of clustering analysis. However, it is unclear

if using the same set of parameters to analyze the count matrices from exonic mappings and mappings against the optimized reference is optimal, since the optimized reference contains more reads and, therefore, potentially more information. We suggest the authors perform a bit more parameter exploration for exonically mapped data, including increasing # of variable features, PCs, k in kNN, and the resolution parameter in the clustering algorithm. If the clusters found in optimal reference cannot be found by using this expanded parameter set, it will be more convincing that the optimal reference itself is critical in discovering the clusters that could not be discovered when using an exonic reference under a reasonable parameter sweep. As for the newly discovered differentially expressed genes, we suggest the authors use a slightly bigger list for exonically mapped data, because it is possible that some of the top DEGs in optimized reference data are not the *top DEGs* in exonically mapped data, but nonetheless show non-trivial fold change and p-value. If this is the case, we would suggest mentioning this in the text and describe that the optimized reference helps emphasize those important genes (i.e. drawing them to the top of the DEG list), which might be missed in the exonically mapped data result as they are not among the top DEGs.

12. In line 318, change any given experiment to all tested experiments.

Typos:

Line 65: we suggest changing to ", and (iii) duplicates are removed"
Line 66: There is nothing after (iv)
Line 32: space around dash "-" are inconsistent.
Line 168: quotes of "premature start transcripts"
Line 221: missing % after "30-40"
Line 223: Extended Fig. 5a doesn't exist
Line 464: space after parentheses
Line 689: commonly detected by the latter. It is unclear what "the latter" means
Line 694: quotes around "regular pre-mRNA annotation" is in the wrong format
Line 700: quote after "(ii)" is in wrong format
In extended data figure 2 (a), the annotation of black bar should be shared genes, instead of shared reads
In the paper, file names are sometimes wrapped by "<>", sometimes by single quotes, and sometimes by double quotes.

Decision Letter, fifth revision:

Response to reviewer comments – June 2023

Reviewer #1:
Remarks to the Author:
The have authors have adequately addressed each of my questions in their revised manuscript.

Thank you!

Reviewer #2:
Remarks to the Author:
We thank the authors for addressing all of the comments and suggestions and answering all of the questions we raised in the review.

Overall, from the rebuttals and updated manuscript, it is clear that all three improvements of the optimized reference over traditional spliced transcriptome reference contribute to the increased gene detection rate and the newly discovered cell markers.

Questions and suggestions:

1. In the response to reviewers, the authors said that they ran STARsolo with the EM algorithm, but in the method section, we did not see --soloMultiMappers EM anywhere. We suggest the authors make explicit mention of how STARsolo was run with the EM, and update the command listed in the methods section to include this option, as the result was discussed thoroughly in the manuscript.

Now added to the "Sequencing, read-mapping and generation of digital expression data" section in Methods.

2. It is unclear specifically what the "intronic reference" refers to in Fig 5(b) title. Please clarify it in the caption.

Amended figure 5a and 5b, where "exon + intron reference" in both figure panels refers to the same thing (Cell Ranger –include-introns mode mapped data) and added the explanation to figure caption.

3. In Ext Fig 3(a), we suggest changing the color of the bar that represents "regular intronic" from black to another color, as black seems to be assigned multiple meanings in the same plot.

Fixed, changed to yellow.

4. In Ext Fig 2 and line 178, It is unclear why adding a transcript length exon (hybrid pre-mRNA reference) for each transcript improves the result of –include-introns. It will be useful if the authors can briefly explain this somewhere in the text, as the difference of this approach from the standard "geneFull" -soloFeatures flag is not clear.

We now elaborate on the reasoning under the revised heading “Reference Transcriptome Optimization“. We essentially found a group of genes with clear intronic read mapping that were not detected by other methods requiring the addition of gene length exons.

5. In Ext Fig 2 (a), the leftmost label on the x-axis, “Regular pre-mRNA reference”, should be changed to “Cell Ranger pre-mRNA reference”.

Fixed.

6. In Figure 1(f), it is obvious that the newly detected genes are found by adding introns and extending 3’ UTRs overlap. Still, the difference is difficult to see without the exact number of yellow and pink bars. Therefore, we suggest the authors add the number represented by these two bars, just like the other two in the same plot.

Done.

7. In line 24: It is unclear if each of these issues impacts the detection of thousands of genes from Extended Fig 3.

The estimates of how many genes are impacted by characterized issues with sequencing read incorporation are based on Fig 1f for intronic read and 3’ intergenic read estimates, and on Fig 3a on gene overlap problems (2000-5000 same-strand overlapping genes depending on species). Extended Data Figure 3 just shows the improvement for two individual datasets which are not expected to reflect the full scope of problems for the entire species as not all genes are expressed in a given tissue.

8. It is unclear how the terminology “readthrough transcript” is used in the manuscript, and it is not exactly the same as what is introduced in an Ensembl blog (https://www.ensembl.info/2019/02/11/annotating-readthrough-transcription-in-ensembl/). We suggest the authors clarify this in the text.

We are precisely addressing the issue outlined in this ensemble blog post. We have also explicitly outlined what we deem as readthrough transcripts in line 175 of our previously submitted manuscript (note that the line numbering for our submitted manuscript and what the reviewers are using appears to be a bit different).

9. In line 304, reference #41 only compared STARsolo and Kallisto but not other methods. The conclusion from that paper was STARsolo outperforms Kallisto w.r.t. detecting genes, memory usage, and # of mapped reads. It is unclear how that conclusion is related to the point discussed in the current manuscript.

We now added another reference that comprehensively compares read integration efficiency between 10 different scRNA-seq analysis pipelines/software and specifically analyzes the differences in intronic

read incorporation. Furthermore, we also now refer the reader to Extended Data Figure 2 in our paper where we directly compare the efficiency of different aligners with respect to generating gene expression estimates.

10. In the "Recovering intronic reads" section (Line 678), the authors introduced three ways to build an intronic reference. The traditional approach, the approach adopted by STARsolo and Cell Ranger, and the new hybrid approach. It is unclear which one was compared in Figure 5 (a) and (b) and Line 306 in the Discussion section and was called "regular exonic and intronic read incorporating reference". We suggest the authors clarify it somewhere in the text or in the caption of Figure 5.

We have now clarified it in Figure 5 caption.

11. In the section Analysis of MnPO and PBMC scRNA-seq data for benchmarking optimized reference performance, the authors clearly explained the procedure of clustering analysis. However, it is unclear if using the same set of parameters to analyze the count matrices from exonic mappings and mappings against the optimized reference is optimal, since the optimized reference contains more reads and, therefore, potentially more information. We suggest the authors perform a bit more parameter exploration for exonically mapped data, including increasing # of variable features, PCs, k in kNN, and the resolution parameter in the clustering algorithm. If the clusters found in optimal reference cannot be found by using this expanded parameter set, it will be more convincing that the optimal reference itself is critical in discovering the clusters that could not be discovered when using an exonic reference under a reasonable parameter sweep. As for the newly discovered differentially expressed genes, we suggest the authors use a slightly bigger list for exonically mapped data, because it is possible that some of the top DEGs in optimized reference data are not the *top DEGs* in exonically mapped data, but nonetheless show non-trivial fold change and p-value. If this is the case, we would suggest mentioning this in the text and describe that the optimized reference helps emphasize those important genes (i.e. drawing them to the top of the DEG list), which might be missed in the exonically mapped data result as they are not among the top DEGs.

We performed extensive parameter exploration with respect to # of variable genes, # of principal components and community discovery resolution parameter for both MnPO and PBMC datasets. The reason why some cell-types are not detected with exonically mapped data vary depending on the precise cell-type. There are some cell types (e.g. T-regs in the PBMC dataset) that eventually cluster distinctly from the rest of the cell types upon increased clustering resolution, but require substantial over-clustering that splits other uniform cell classes to many artifactual subclusters. There are yet other cell-types (e.g. MnPO-Glut3 identified with optimal reference genome in Extended Data Figure 4) that would not be identified at all since the defining genes are not in the dataset. Finally, we also see cell-types that can easily be distinguished as a separate cell type with an optimized reference but do not appear as individual neuron types at most parameters tested due to lack of expression information from genes that delineate these clusters (e.g. MAIT cells and CD8 T terminal effectors) with suboptimally retrieved expression information. The bottom line is, that optimal recovery of the full gene expression data with optimized

references makes the cell-type discovery more reliable in a wider parameter space with the outcome substantially more likely to reflect biological reality. We added a sentence in discussion to clarify that.

12. In line 318, change any given experiment to all tested experiments. ➔ done

Typos:

Line 65: we suggest changing to ", and (iii) duplicates are removed" ➔ done
Line 66: There is nothing after (iv) ➔ fixed, the location of numeration was previously confusing
Line 32: space around dash "-" are inconsistent. ➔ fixed
Line 168: quotes of "premature start transcripts" ➔ fixed
Line 221: missing % after "30-40" ➔ fixed
Line 223: Extended Fig. 5a doesn't exist ➔ Fixed to 3a
Line 464: space after parentheses ➔ fixed
Line 689: commonly detected by the latter. It is unclear what "the latter" means ➔ fixed
Line 694: quotes around "regular pre-mRNA annotation" is in the wrong format ➔ fixed
Line 700: quote after "(ii)" is in wrong format ➔ fixed
In extended data figure 2 (a), the annotation of black bar should be shared genes, instead of shared reads ➔ fixed
In the paper, file names are sometimes wrapped by "<>", sometimes by single quotes, and sometimes by double quotes. ➔ changed everything uniformly to ""

Final Decision Letter:

15th Aug 2023

Dear Dr Oka,

I am pleased to inform you that your Article, "Recovery of missing single-cell RNA-sequencing data with optimized transcriptomic references", has now been accepted for publication in Nature Methods. Your paper is tentatively scheduled for publication in our November print issue, and will be published online prior to that. The received and accepted dates will be 20th Apr 2022 and 15th Aug 2023. This note is intended to let you know what to expect from us over the next month or so, and to let you know where to address any further questions.

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.

Over the next few weeks, your paper will be copyedited to ensure that it conforms to Nature Methods 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.

Please note that Nature Methods 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 self-archiving policies. Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

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

Your paper will now be copyedited to ensure that it conforms to Nature Methods style. Once proofs are generated, they will be sent to you electronically and you will be asked to send a corrected version within 24 hours. It is extremely important that you let us know now whether you will be difficult to contact over the next month. If this is the case, we ask that you send us the contact information (email, phone and fax) of someone who will be able to check the proofs and deal with any last-minute problems.

Once your manuscript is typeset and you have completed the appropriate grant of rights, 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.

Once your paper has been scheduled for online publication, the Nature press office will be in touch to confirm the details.

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.

Content is published online weekly on Mondays and Thursdays, and the embargo is set at 16:00 London time (GMT)/11:00 am US Eastern time (EST) on the day of publication. If you need to know the exact publication date or when the news embargo will be lifted, please contact our press office after you have submitted your 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 NMETH-A46722E and the name of the journal, which they will need when they contact our office.

About one week before your paper is published online, we shall be distributing a press release to news organizations worldwide, which may 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 Methods. Our Press Office will contact you closer to the time of publication, but if you or your Press Office have any inquiries in the meantime, please contact press@nature.com.

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.

Nature Portfolio journals encourage authors to share their step-by-step experimental protocols on a protocol sharing platform of their choice. Nature Portfolio 's Protocol Exchange is a free-to-use and open resource for protocols; protocols deposited in Protocol Exchange are citable and can be linked from the published article. More details can found at www.nature.com/protocolexchange/about.

Please note that you and any of your coauthors will be able to order reprints and single copies of the issue containing your article through Nature Portfolio's reprint website, which is located at http://www.nature.com/reprints/author-reprints.html. If there are any questions about reprints please send an email to author-reprints@nature.com and someone will assist you.

Please feel free to contact me if you have questions about any of these points. Thank you very much for publishing your paper at Nature Methods!

Best regards,

Lin Tang, PhD
Senior Editor
Nature Methods