Cell Systems Transparent Peer Review Record

Nonparametric single-cell multiomic characterization of trio relationships between transcription factors, target genes, and cis-regulatory regions

Yuchao Jiang,Yuriko Harigaya, Zhaojun Zhang, Hongpan Zhang, Chongzhi Zang, Nancy R. Zhang

Summary

Initial Submission:	Received April 22, 2022 Preprint: https://doi.org/10.1101/2021.09.22.461437 Scientific editor: Ernesto Andrianantoandro, Ph.D.
First round of review:	Number of reviewers: Two *Two confidential, Zero signed* Revision invited June 06, 2022 *Major changes anticipated* Revision received June 23, 2022
Second round of review:	Number of reviewers: Two *Two original, Zero new* *Two confidential, Zero signed* Accepted August 11, 2022

CellPress

This Transparent Peer Review Record is not systematically proofread, type-set, or edited. Special characters, formatting, and equations may fail to render properly. Standard procedural text within the editor's letters has been deleted for the sake of brevity, but all official correspondence specific to the manuscript has been preserved.

Editorial decision letter with reviewers' comments, first round of review

Dear Dr. Jiang,

I hope this email finds you well. I do appreciate that this manuscript has had an extensive history at another publisher. I believe, in the case of this manuscript, it is best to have a blank slate for evaluation and review at Cell Systems. I have elected to treat this as a fresh submission and recruited reviewers accordingly. The reviews of your manuscript are back and I've appended them below. On balance, the reviewers appreciate the goals of the work presented here; they've provided constructive comments that are aligned with our hopes for the paper. Accordingly, we're happy to invite a revision.

To help guide this revision, here are a few points that I see as being potentially make-or-break as we move forward:

1. The description of the methodology, and justifications for choices made, needs to be clarified.
2. The argument for utility (both the advantage over other approaches and the capacity for enabling biological insight) needs to be better articulated and substantiated.

In addition to these points, I've highlighted reviewer comments that seem to warrant special attention. I'd also like to be explicit about an almost philosophical stance that we take at Cell Systems:

We believe that understanding how approaches fail is fundamentally interesting: it provides critical insight into understanding how they work. We also believe that all approaches do fail and that it's unreasonable, even misleading, to expect otherwise. Accordingly, when papers are transparent and forthright about the limitations and crucial contingencies of their approaches, we consider that to be a great strength, not a weakness. Please keep this in mind when addressing the reviewer concerns.

I hope you find this feedback helpful. If you have any questions or concerns, I'm always happy to talk, either over email or by Zoom. More technical information and advice about resubmission can be found below my signature. Please read it carefully, as it can save substantial time and effort later.

I look forward to seeing your revised manuscript.

All the best,

Ernesto Andrianantoandro, Ph.D.
Scientific Editor, Cell Systems

Reviewers' comments:

Reviewer #1: Jiang et al. presented a nonparametric method TRIPOD to discover regulatory trios consisting of genes, accessible peaks, and TFs in single-cell multi-omics data (RNA-seq and ATAC-seq). The idea is useful and it is a necessary molecular detail that we want to know from single-cell data. The authors analyzed three datasets and presented interesting examples showing the method's sensitivity. However, we have major concerns about the presentation of the new method, several parts of the analyses, and the molecular-level assumptions. The comments are given below.

Presentation of TRIPOD:

1. TRIPOD has been presented to increase "robustness and scalability" compared to PECA's parametric model. However, there is no comparison on scalability or robustness.
2. In fact, the authors seemed to push too hard against parametric models without really demonstrating the added benefit. An explanation like "The main motivation for using the matching model above is our reluctance to assume the simple linear relationship" is not object. The point is not about being reluctant. The method should show examples where parametric linear models fail to conform to biological insight and the non-parametric formulation recovers them.
3. Along the same line, the authors mention that "The estimated coefficients from the nonparametric and parametric methods are correlated on the global scale (Supplementary Fig. 13 ..." but S13 doesn't show the correlation between the \alpha and the \beta parameters. Since S13 doesn't show this information, one needs to rely on S27 and S28, but these figures present only four genes and the key claim ("The estimated coefficients from the … are correlated on the global scale") remains poorly demonstrated.
4. Overall, S13 is difficult to read. Side comment: it is unclear if gene~peak marginal not correlating well with match TF level 2 or gene~TF marginal not correlating with match peak level 1 is intuitive.
5. Influence analysis - presented in the linear regression framework, not clear if the authors actually switched to using a linear model in this part or used linear regression just for the ease of presentation.

Different parts of the analyses:

6. metacell: the authors did not provide the details of computing metacells. Knowing the resolution of metacells is important. It seems the authors introduced the name - they cited Hao et al. Cell 2021 but that paper doesn't use the term. However, Amos Tanay has used it and defined it more rigorously in their Genome Biology 2019 paper. The authors need to clarify their operations.
7. How much variability is introduced because of the metacell sizes? Fig 2 suggests the gene expression in a metacell is not normalized for the size.
8. It is also unclear why level 2 tests average the TF expression? metacell matching is not on a gene by gene basis, but on the basis of the whole transcriptome (or a fixed number of PCs?), so it is unclear clear how much variability is lost due to averaging.

9. "Note that even though TF expression is not included in the model as a main term, it is controlled for (and not just in the linear sense) by the matching." This point and the rest of the paragraph in the Method section are unclear.
10. It is difficult to contrast the two visualizations in Figs S5 and 2.
11. In the FGL2-MAFK analysis, the authors note that "Both observations suggest that … we do not claim that the significantly linked peaks and TFs through marginal association are false positives, but rather this specific trio is insignificant …." This comment is acceptable but does not show the true complementary value of TRIPOD. What would be important here is to show that a marginal test discovers an association and TRIPOD finds it as a false positive or pinpoints the peak/TF underlying that association.
12. "Additionally, the unique sets of trio regulatory relationships identified by TRIPOD but not by random matching … have significant enrichment, demonstrating the effectiveness of TRIPOD in identifying true trio relationships … (Supplementary Fig. 14b)." difficult to assess this point in S14b since the scale of the Y-axis is not logarithmic.
13. "Importantly, TRIPOD detected a substantial number of peak-TF pairs that were not detected through marginal associations but validated by ChIP-seq (Fig. 5b)." - is this true for Nurog2 and Eomes?
14. "we established a link if at least one of the TRIPOD tests returned a positive coefficient estimate with FDR-adjusted p-values less than 0.01 for at least one trio involving the pair of the TF and the target gene" - this sounds very permissive and it is important to know given this much permissibility, did TRIPOD discover any novel link (hitherto unknown)? Also, for future studies, is this much permissibility recommended?

Molecular assumptions

15. TRIPOD assumes that the mRNA expression of a TF is a proxy for its protein expression. This is generally not true. Given current data availability, we understand that the authors will need to make this assumption but it is also important to (a) discuss this point and (b) whenever possible, check if protein expression data is available for the TFs discovered in TRIPOD trios (human protein atlas?) and test if this assumption holds for them.

Reviewer #2: The authors here present TRIPOD which is an alternative framework to test the regulatory relations between peaks and gene expression in multi-omic data sets (Figure 2). TRIPOD uses target gene expression, chromatin accessibility of motif containing peaks, and expression of the motif linked transcription factor as 3 variables. Of these 2 are used to in an either-or fashion (always including target gene expression) to establish a regulatory relationship. This relationship is strengthened or weakened using the link with the 3rd variable that wasn't used (level 1). In level 2 analysis the relationship between 2 variables becomes conditional upon the unused relationship and as such more stringent. They show examples of how false positives and false negatives can be corrected for using this method (Figure 2). They benchmark their method to marginal (using 2 variables) analyses and show that in most cases, albeit not all, using a variety of published datasets the analysis is improved (Figures 3-6). In addition, an unexpected link was found between four neurogenesis transcription factors and gliogenic target genes (Figure 5) although little follow up was done.

While others have also used motifs, accessibility and target gene expression, the current method also factors in transcription factor expression which is new (Figure 1). Nevertheless, it's understandable that adding additional variables makes the analysis of regulatory relationships more stringent but generally this comes at a cost of sensitivity. The authors acknowledge this and discuss this handily in the discussion section suggesting it is not the point of the analysis to find everything. However, in using the level 1 analysis I wonder whether much is lost in terms of sensitivity as the analysis works in and either-or fashion adding what is found to the pool of putative regulatory relationships. While no analysis is be perfect, overall, I'm enthusiastic about the approach the authors take and I believe that the strategy the authors use (especially strengthening or weakening relationships using the unused variable) represents an improvement of what is currently being used. Also, the ability to choose the level of analysis is I think an elegant way to deal with sensitivity concerns.

I have several comments the authors could address.
* The author mentioned in manuscript lines 128:130 "Only peaks within a certain range of the genen's TSS and only interaction between Tfs and peaks containing high-scoring binding motifs for that Tfs are considered". It is unclear to me how these threshold/cutoffs are calculated/determined.
* The author mentioned in manuscript lines 134:135 "To avoid overfitting, we performed out-of-fold prediction and adopted independent training and testing sets". Can the author describe the method more clearly? For example if the author used random forest it should be clear what parameters are used and what the training and test sets are.
* Lines 141:142 "To mitigate the consequences of sparsity …., we clustered cells to form metacells". Is this nearest neighbor clustering (Seurat)? If so, please describe briefly how it is done. I also assume that this is done for the all-benchmark data sets the same way.
* Lines 156:157 are describing how the pipeline deals with LASSO selection behavior. It is known that LASSO selects a maximum of n variables before saturation. How do they deal with selecting the n and also with a group of variables with high pairwise correlation.
* Lines 191:199. As I understand correctly for each metacell-pair a dxp, dyt and dyg is calculated, then metacell-pairs that have low values for each of these measurements, are masked. If this is true, how is this low value determined and to what extend does this method affect the power of the data regarding making it more sparse.

Authors' response to the reviewers' first round comments

Attached.

Editorial decision letter with reviewers' comments, second round of review

Dear Dr. Jiang,

I'm very pleased to let you know that the reviews of your revised manuscript are back, the peer-review process is complete, and only a few minor, editorially-guided changes are needed to move forward towards publication.

In addition to the final comments from the reviewers, I've made some suggestions about your manuscript within the "Editorial Notes" section, below. Please consider my editorial suggestions carefully, ask any questions of me that you need, make all warranted changes, and then upload your final files into Editorial Manager.

I'm looking forward to going through these last steps with you. Although we ask that our editorially-guided changes be your primary focus for the moment, you may wish to consult our FAQ (final formatting checks tab) to make the final steps to publication go more smoothly. More technical information can be found below my signature, and please let me know if you have any questions.

All the best,

Ernesto Andrianantoandro, Ph.D.
Scientific Editor, Cell Systems

Editorial Notes

Transparent Peer Review: Thank you for electing to make your manuscript's peer review process transparent. As part of our approach to Transparent Peer Review, we ask that you add the following sentence to the end of your abstract: "A record of this paper's Transparent Peer Review process is included in the Supplemental Information." Note that this ***doesn't*** count towards your 150 word total!

Also, if you've deposited your work on a preprint server, that's great! Please drop me a quick email with your preprint's DOI and I'll make sure it's properly credited within your Transparent Peer Review record.

Title: Your title does not fully convey the conceptual advance of the paper. I suspect it could be more effective. Please include something about the three-way relationship between TFs, their target genes, and chromatin accessibility. For example: “Nonparametric single-cell omics characterization of three-way relationships between transcription factors, target genes, and chromatin accessibility”

As you re-consider your title, note that an effective title is easily found on Pubmed and Google. A trick for thinking about titles is this: ask yourself, "How would I structure a Pubmed search to find this paper?" Put that search together and see whether it comes up is good "sister literature" for this work. If it does, feature the search terms in your title. You also may wish to consider that PubMed is sensitive to small differences in search terms. For example, “NF-kappaB” returned ~84k hits as of March, 2018, whereas “NFkappaB” only returned ~8200. Please ensure that your title contains the most effective version of the search terms you feature.

Abstract: The abstract reads wonderfully, but is unfortunately too long. Please condense to 150 words or less.

Manuscript Text:

- House style disallows editorializing within the text (e.g. strikingly, surprisingly, importantly, etc.), especially the Results section. These terms are a distraction and they aren't needed—your excellent observations are certainly impactful enough to stand on their own. Please remove these words and others like them. “Notably” is suitably neutral to use once or twice if absolutely necessary.
- We don’t allow “priority claims” (e.g. new, novel, etc.). For a discussion of why, read: http://crosstalk.cell.com/blog/getting-priorities-right-with-novelty-claims, http://crosstalk.cell.com/blog/novel-insights-into-priority-claims.
- There can only be one reference list - please incorporate the references from the supplement into the main reference list.

Figures and Legends:

Please look over your figures keeping the following in mind:

- When data visualization tools are used (e.g. UMAP, tSNE), please ensure that the dataset being visualized is named in the figure legend and, when applicable, its accession number is included.

- When color scales are used, please define them, noting units or indicating "arbitrary units," and specify whether the scale is linear or log.
- Please ensure that every time you have used a graph, you have defined "n's" specifically and listed statistical tests within your figure legend.
- Please ensure that all figures included in your point-by-point response to the reviewers' comments are present within the final version of the paper, either within the main text or within the Supplemental Information.

STAR Methods: Note that Cell Press has recently changed the way it approaches "availability" statements for the sake of ease and clarity. Please revise the first section of your STAR Methods as follows, noting that the particular examples used might not pertain to your study. Please consult the STAR Methods guidelines for additional information.

RESOURCE AVAILABILITY

Lead Contact: Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Jane Doe (janedoe@qwerty.com).

Materials Availability: This study did not generate new materials. -*OR*- Plasmids generated in this study have been deposited at [Addgene, name and catalog number]. -*OR*- etc.

Data and Code Availability:

- **Source data statement** (described below)
- **Code statement** (described below)
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Data and Code Availability statements **have three parts and each part must be present. Each part should be listed as a bullet point, as indicated above.**

Instructions for section 1: Data. The statements below may be used in any number or combination, but at least one must be present. They can be edited to suit your circumstance. Please ensure that all datatypes reported in your paper are represented in section 1. For more information, please consult this list of standardized datatypes and repositories recommended by Cell Press.

• [Standardized datatype] data have been deposited at [datatype-specific repository] and are publicly available as of the date of publication. Accession numbers are listed in the key resources table.

- [Adjective] data have been deposited at [general-purpose repository] and are publicly available as of the date of publication. DOIs are listed in the key resources table.

- [De-identified human/patient standardized datatype] data have been deposited at [datatype-specific repository]. They are publicly available as of the date of publication until [date or delete “until”]. Accession numbers are listed in the key resources table.

- [De-identified human/patient standardized datatype] data have been deposited at [datatype-specific repository], and accession numbers are listed in the key resources table. They are available upon request until [date or delete “until”] if access is granted. To request access, contact [insert name of governing body and instructions for requesting access]. [Insert the following when applicable] In addition, [summary statistics describing these data/processed datasets derived from these data] have been deposited at [datatype-specific repository] and are publicly available as of the date of publication. These accession numbers are also listed in the key resources table.

- Raw [standardized datatype] data derived from human samples have been deposited at [datatype-specific repository], and accession numbers are listed in the key resources table. Local law prohibits depositing raw [standardized datatype] datasets derived from human samples outside of the country of origin. Prior to publication, the authors officially requested that the raw [adjective] datasets reported in this paper be made publicly accessible. To request access, contact [insert name of governing body and instructions for requesting access]. [Insert the following when applicable] In addition, [summary statistics describing these data/processed datasets derived from these data] have been deposited at [datatype-specific repository] and are publicly available as of the date of publication. These accession numbers are also listed in the key resources table.

- The [adjective] data reported in this study cannot be deposited in a public repository because [reason]. To request access, contact [insert name of governing body and instructions for requesting access]. [Insert the following when applicable] In addition, [summary statistics describing these data/processed datasets derived from these data] have been deposited at [datatype-specific or general-purpose repository] and are publicly available as of the date of publication. [Accession numbers or DOIs] are listed in the key resources table.

- This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.

- [Adjective or all] data reported in this paper will be shared by the lead contact upon request.

Instructions for section 2: Code. The statements below may be used in any number or combination, but at least one must be present. They can be edited to suit your circumstance. ***If you are using GitHub, please follow the instructions here to archive a “version of record” of your GitHub repo at Zenodo, then report the resulting DOI. Additionally, please note that the Cell Systems strongly recommends that you also include an explicit reference to any scripts you may have used throughout your analysis or to generate your figures within section 2.***

• All original code has been deposited at [repository] and is publicly available as of the date of publication. DOIs are listed in the key resources table.

• All original code is available in this paper's supplemental information.

• This paper does not report original code.

Instructions for section 3. Section 3 consists of the following statement: Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

In addition,

- STAR Methods follows a standardized structure. Please reorganize your experimental procedures to include these specific headings in the following order: LEAD CONTACT AND MATERIALS AVAILABILITY (including the three statements detailed above); EXPERIMENTAL MODEL AND SUBJECT DETAILS (when appropriate)*;* METHOD DETAILS (required); QUANTIFICATION AND STATISTICAL ANALYSIS (when appropriate); ADDITIONAL RESOURCES (when appropriate). We're happy to be flexible about how each section is organized and encourage useful subheadings, but the required sections need to be there, with their headings. They should also be in the order listed. Please see the STAR Methods guide for more information or contact me for help.
- Please ensure that original code has been archived in a general purpose repository recommended by Cell Press and that its DOI is provided in the Software and Algorithms section of the Key Resources Table. If you've chosen to use GitHub, please follow the instructions here to archive a "version of record" of your GitHub repo at Zenodo, complete with a DOI.
- Currently, you don't have a **Key Resources Table** (KRT). Note that the key resources table is required for manuscripts with an experimental component, and if a purely computational manuscript links to any external datasets (previously published or new), code-containing websites (e.g. a GitHub repo, noting that DOIs are strongly preferred), or uses non-standard software, it needs to include a key resources table that details these aspects of the paper. Purely computational or theoretical papers that don't contain any external links and use standard software don't require a key resources table, although you're welcome to include one if you like. For details, please refer to the Table Template or feel free to ask me for help.

Thank you!

Reviewer comments:

Reviewer #1: < none >

Reviewer #2: I have no further questions for the authors and recommend publication.

We thank the editor and the reviewers for their encouraging and constructive feedback. Below are our point-by-point responses to the reviewers' comments. The original reviewers' comments are in black with the editor's highlights, and our responses are in blue. The changes that we made are highlighted in the main manuscript and supplements.

Reviewer #1

Jiang et al. presented a nonparametric method TRIPOD to discover regulatory trios consisting of genes, accessible peaks, and TFs in single-cell multi-omics data (RNA-seq and ATAC-seq). The idea is useful, and it is a necessary molecular detail that we want to know from single-cell data. The authors analyzed three datasets and presented interesting examples showing the method's sensitivity. However, we have major concerns about the presentation of the new method, several parts of the analyses, and the molecular-level assumptions. The comments are given below.

Presentation of TRIPOD:

1. TRIPOD has been presented to increase "robustness and scalability" compared to PECA's parametric model. However, there is no comparison on scalability or robustness.

PECA is developed for bulk gene expression and chromatin accessibility data and uses pairings of these two modalities to infer gene regulatory networks. The software (https://github.com/SUwonglab/PECA) requires three input files: a tab delimited gene expression file (two columns, gene symbol and expression value), a bam file containing the chromatin accessibility data from DNase-seq or ATAC-seq, and the index file of the bam file (.bai). As such, PECA runs on a single pair of RNA-seq and DNase-seq/ATAC-seq data and cannot be run on the single-cell multiomics data.

We have rewritten the reference to PECA so as to clarify that although it is a multimodal model for joint analysis of bulk ATAC and bulk RNA data, it is not meant for (and does not accept as input) single cell multiome data. The revision is as follows:

"One exception is PECA (Duren et al., 2017), which uses a parametric model to characterize the joint four-way relationship between TF expression, regulatory site accessibility, chromatin remodeler expression, and target gene expression. PECA was designed for matched bulk transcriptomic and epigenomic data and does not work for single cell multiome data. In this paper, we propose a scalable nonparametric approach for joint modeling of single cell multiome RNA and ATAC sequencing data."

2. In fact, the authors seemed to push too hard against parametric models without really demonstrating the added benefit. An explanation like "The main motivation for using the matching model above is our reluctance to assume the simple linear relationship" is not object. The point is not about being reluctant. The method should show examples where parametric linear models fail to conform to biological insight and the nonparametric formulation recovers them.

We have reworded the corresponding sentence to "The main motivation for using the matching model above is to avoid making linearity assumptions on the relationship between any of the variables."

We have made very extensive benchmark analyses between TRIPOD's nonparametric approach and the linear model, with results shown in Supplementary Fig. 23, 24, and 25. While the estimated coefficients from the nonparametric and parametric methods are correlated on the global scale, they are different as can be seen zooming into specific genes (Supplementary Fig. 24). (1) The nonparametric model of TRIPOD is more stringent (Supplementary Fig. 25) by controlling for not only linear conditional relationships but also nonlinear ones. (2) The nonparametric approach is more robust to outliers with the use of the rank-based Spearman correlation. Therefore, the significant call sets from the parametric and nonparametric overlap but are not exactly the same.

Importantly, we do not think that this parametric v.s. nonparametric difference is the only contribution of TRIPOD, in that linear models involving three factors with and without interactions have also not been explored for the identification of regulatory relationships between transcription factor, *cis* region, and gene expression. Since the parametric form of such interactions is rarely known, we chose the nonparametric approach implemented in TRIPOD, which avoids the stringent assumptions of the linear relationships and the Gaussian errors. We demonstrate the effectiveness of TRIPOD's nonparametric approach with all the extensive validation and benchmark analyses carried out.

3. Along the same line, the authors mention that "The estimated coefficients from the nonparametric and parametric methods are correlated on the global scale (Supplementary Fig. 13 ..." but S13 doesn't show the correlation between the \alpha and the \beta parameters. Since S13 doesn't show this information, one needs to rely on S27 and S28, but these figures present only four genes and the key claim ("The estimated coefficients from the … are correlated on the global scale") remains poorly demonstrated.

Thanks. In this revision, we generated the pairwise correlation plots of the coefficients on the global scale in the re-numbered Supplementary Fig. 24e. The key claim as mentioned by the reviewer generalizes from the four genes to the genome-wide scale.

4. Overall, S13 is difficult to read. Side comment: it is unclear if gene~peak marginal not correlating well with match TF level 2 or gene~TF marginal not correlating with match peak level 2 is intuitive.

We don't expect the detections made by marginal models to correlate well with the level 1 and level 2 detections made by TRIPOD, because TRIPOD is interrogating conditional relationships. Thanks for your question as they made us realize that this point is not clear and should be emphasized. In the revision, we have rewritten many parts of the paper to clarify this point. In the introduction:

"We show that the detections are coherent with the vast amounts of existing knowledge from enhancer databases, bulk cell-type-specific chromatin immunoprecipitation followed by sequencing (ChIP-seq) experiments, tissue-specific TF knockdown/knockout studies, and cis-eQTL studies, but that conditional and marginal models identify different sets of relationships. This is expected, as regulatory relationships are complex, and conditional and marginal models are interrogating different types of associations in the data."

In the results:

"Existing methods (Hao et al., 2021; Ma et al., 2020; Stuart et al., 2021) screen for marginal associations either between the TF and the peak or between the peak and the target gene. However, three-way relationships may be complex: When a TF binds to a cis-regulatory region to affect the expression of a gene, it can do so in multiple ways, leading to different patterns in the data. The TF could be directly responsible for opening the chromatin of the enhancer region, facilitating the binding of other TFs that recruit the RNA polymerase. In such cases, expression of the TF is likely to be marginally correlated with the accessibility of the enhancer region, but its correlation with the expression of the target gene may be masked due to confounding of other involved TFs. Alternatively, the TF may not be directly responsible for chromatin remodeling but may bind to already accessible chromatin in recruiting other TFs or the RNA polymerase. In such cases, expression availability of the TF may not be highly correlated with the accessibility of the enhancer region, but conditioned on the accessibility of the enhancer region, it should be associated with the expression of the target gene. When marginal associations are masked, evidence for binding of the TF at the peak in the regulation of a gene can be inferred from partial associations: (i) with the peak open at a fixed accessibility, whether cells with higher TF expression have higher gene expression; and (ii) with the TF expression fixed at a value above a threshold, whether cells with higher peak accessibility have higher gene expression."

However, we feel that Supplementary Figure 13 is still informative for visualizing the degree of concordance between TRIPOD estimates of conditional associations and marginal associations. In referencing it, we added a reminder that the models underlying the different tests are different, and thus we don't expect complete agreement. Figure 13 is merely a birds-eye-view of the degree of similarity between the tests. We updated the caption and legend for Supplementary Fig. 12 (renumbered). Gene~peak marginal correlates with gene~peak matching TF level 1, but NOT level 2. Similarly, gene~TF marginal correlates with gene~TF matching TF level 1, but NOT level 2. This makes sense: After TRIPOD's matching, for the level 1 testing, we test $\Delta Yg=\alpha\Delta Xp$ and $\Delta Yg=\beta\Delta Yt$ via a Spearman correlation test. Without the non-bipartite matching, this is of the same form as the marginal testing. However, with the matching, we are testing the conditional association, rather than the marginal association. On the other hand, for level 2 testing, we additionally test for the interactions, and therefore the results differ significantly from the marginal testing.

5. Influence analysis - presented in the linear regression framework, not clear if the authors actually switched to using a linear model in this part or used linear regression just for the ease of presentation.

> Currently this influence analysis is only conducted under the linear regression framework, since it involves resampling and thus the linear framework allows much faster computation. Since this influence analysis is only done for trios that pass the FDR threshold, we still enjoy the robustness offered by the distribution-free matching-based screening step. By reverting back to a linear model in this step, we gain interpretability of the parameters. We rewrote the description of the Influence analysis to make this clearer:
>
> *"For significant trios, TRIPOD further carries out a sampling-based influence analysis, where phenotypically contiguous sets of cell aggregates are held out to measure their influence on the estimated coefficients. This influence analysis assumes a linear model for computational speed and scalability. The corresponding cell types/states that lead to significant deviations from the null upon their removal have high influence scores, which can be used to identify cell types/states that drive a regulatory relationship."*

Different parts of the analyses:

6. metacell: the authors did not provide the details of computing metacells. Knowing the resolution of metacells is important. It seems the authors introduced the name - they cited Hao et al. Cell 2021 but that paper doesn't use the term. However, Amos Tanay has used it and defined it more rigorously in their Genome Biology 2019 paper. The authors need to clarify their operations.

> Thanks for pointing this out. We agree that the term "metacell" is confusing and therefore changed it to "cell aggregate" in the revision. We have also added to the revision more details of how the cell aggregates are computed in the methods section. Although careful choice of cell-aggregate resolution is important, we do not require that cell aggregates strictly represent "disjoint and homogenous groups of profiles that could have been resampled from the same cell", as Baran et al. strive for in their metacell method. Our cell-aggregate preprocessing step is merely a way to combat the sparsity of the data, while producing a sizable set of points for the conditional association analysis.
>
> We have carried out a systematic evaluation of how different resolutions/constructions of cell aggregates affect TRIPOD's performance. Specifically, we constructed cell aggregates for the mouse embryonic brain data by the Louvain algorithm using the Seurat package as well as the metacell approach by Baran et al. (Genome Biology 20:206, 2019; PMID: 31604482) and evaluated the model performance based on the PLAC-seq and ChIP-seq data. We observed that Baran's method tended to give metacells that fail to pass the data filtering step prior to TRIPOD and, thus, consider that Louvain clustering is more suited to our approach than the Baran's method. The results obtained by Louvain clustering at various resolutions are now shown in Supplementary Fig. 18 – the validation results were robust to changes in resolutions of the cell aggregates.

7. How much variability is introduced because of the metacell sizes? Fig 2 suggests the gene expression in a metacell is not normalized for the size.

> We did normalize for the cell-aggregate library size before analyses, thus the library size should not confound the analysis. This is also done in Fig. 2.

8. It is also unclear why level 2 tests average the TF expression? metacell matching is not on a gene by gene basis, but on the basis of the whole transcriptome (or a fixed number of PCs?), so it is unclear clear how much variability is lost due to averaging.

> The matching of cell aggregates is indeed on the TF-by-TF (or peak-by-peak) basis. We do this because we don't want to lose too much variability. If we were to match by the whole transcriptome, then there would be little variability left in the gene for the regression.
>
> In level 2 test matching by TF, we want to take the average of the TF expression between the matched pair as a better estimate of the true TF expression of that pair. Then, our goal is to determine whether there is an interaction effect of the TF and the peak (where its motif is found) on the target gene expression. This is why we form the

product of the averaged TF and the change (between the pair) in the peak accessibility, and see whether that predicts the change (between the pair) in target gene expression.

9. "Note that even though TF expression is not included in the model as a main term, it is controlled for (and not just in the linear sense) by the matching." This point and the rest of the paragraph in the Method section are unclear.

Thanks for pointing out this confusing sentence. It has been rephrased in the revision:

"Note that even though TF expression is included in this interaction model only as an interaction term (with ΔX_{jp}) and not as a main term, it is controlled marginally by its use as the matching variable."

10. It is difficult to contrast the two visualizations in Figs S5 and 2.

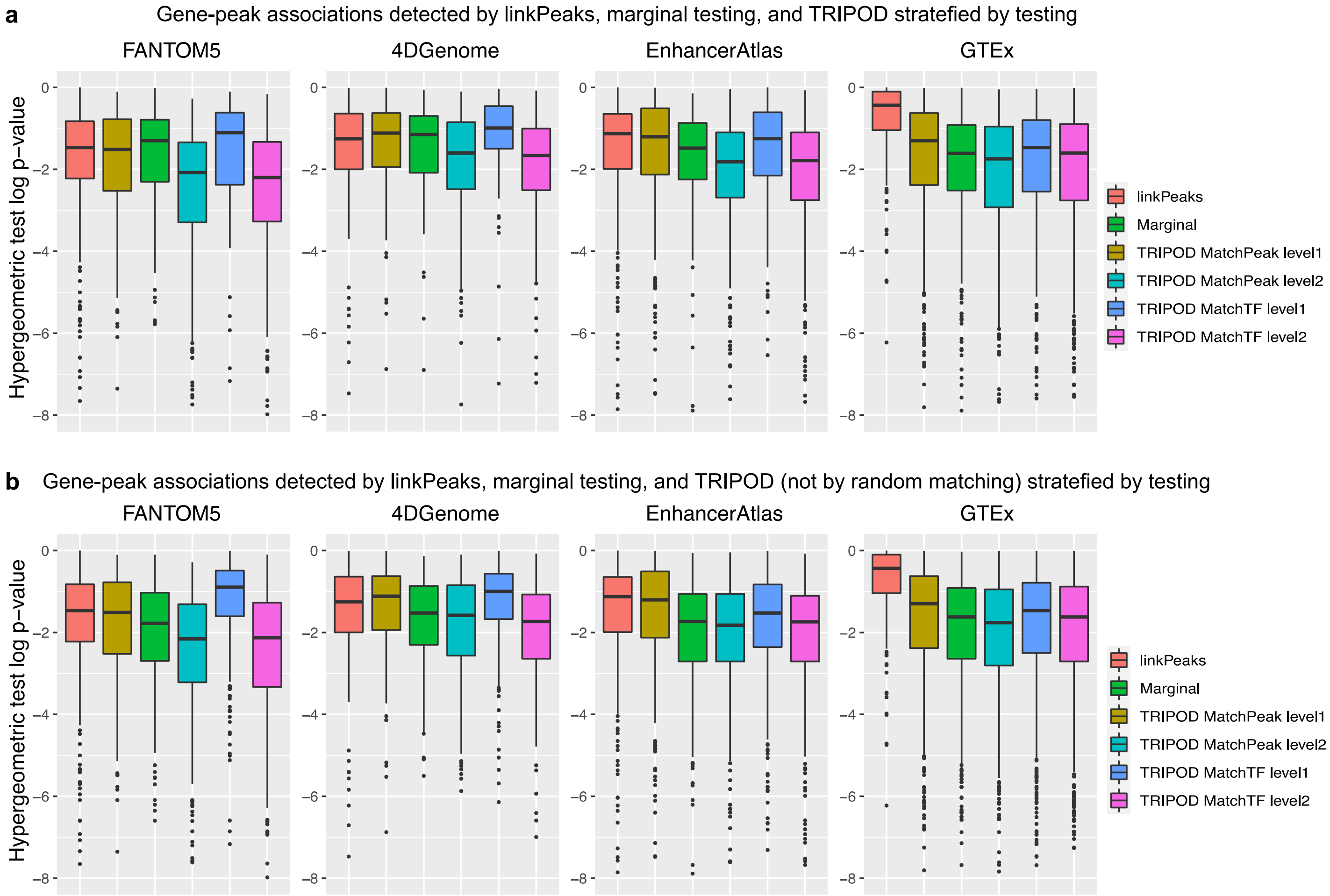
Thanks for pointing this out. We agree that Supplementary Fig. 5 is confusing and does not add much given what is already conveyed in Fig. 2, and thus removed it in the revision.

11. In the FGL2-MAFK analysis, the authors note that "Both observations suggest that … we do not claim that the significantly linked peaks and TFs through marginal association are false positives, but rather this specific trio is insignificant …." This comment is acceptable but does not show the true complementary value of TRIPOD. What would be important here is to show that a marginal test discovers an association and TRIPOD finds it as a false positive or pinpoints the peak/TF underlying that association.

Marginal and conditional association are embodying different relationships. Therefore, we did not purposely focus on identifying false positives from marginal testing, but rather what we wanted to point out is that valid gene-peak and gene-TF marginal relationships do not necessarily lead to a valid gene-peak-TF trio relationship. We have made this point clear in the revised manuscript. Additionally, our paper does contain comprehensive benchmarking where we show that we increase the recovery of known relationships (e.g., precision-recall plots in Fig. 3c,d and the novel trios supported by PLAC-seq and ChIP-seq in Fig. 5a-c).

12. "Additionally, the unique sets of trio regulatory relationships identified by TRIPOD but not by random matching … have significant enrichment, demonstrating the effectiveness of TRIPOD in identifying true trio relationships … (Supplementary Fig. 14b)." difficult to assess this point in S14b since the scale of the Y-axis is not logarithmic.

The main takeaway from the benchmark/validation shown in the renumbered Supplementary Fig. 13 is from comparing panel a (TRIPOD's call sets) to panel b (TRIPOD's call sets minus the ones called by random matching, which is akin to marginal testing) – we show the effectiveness of TRIPOD in identifying additional relationships validated by existing enhancer and *cis*-eQTL database. We replotted the logarithm of the p-values from the hypergeometric test as below, and they do not make much difference. Therefore, we are keeping the p-values untransformed for visualization directly on the original scale.



a Gene-peak associations detected by linkPeaks, marginal testing, and TRIPOD stratefied by testing

b Gene-peak associations detected by linkPeaks, marginal testing, and TRIPOD (not by random matching) stratefied by testing

13. "Importantly, TRIPOD detected a substantial number of peak-TF pairs that were not detected through marginal associations but validated by ChIP-seq (Fig. 5b)." - is this true for Neurog2 and Eomes?

Yes, in Figure 5b, the blue arrows show the regulatory links inferred by TRIPOD that was previously known, and the green arrows show the newly inferred regulatory links that have ChIP-seq validation. We see that some of these links do involve Neurog2 and Eomes: Neurog regulates Sox9 and Nfia, and Eomes regulates Nfia. These links suggest crosstalk between the Neurogenesis and Gliogenesis cascades. They were not previously known, but they are supported by ChIP-seq.

The text is revised to point this out: "*In particular, TRIPOD detected crosstalk between the neurogenesis and gliogenesis regulatory cascades: Neurog2 regulates Sox9 and Nfia, and Eomes and Tbr1 regulates Nfia. The partial residual plots (Fig. 5d) show the strong partial associations in TRIPOD level 1 tests. These new regulatory relationships are supported by ChIP-seq peaks for Neurog2 around Sox9 and Nfia, and for Eomes and Tbr1 around Nfia.*"

14. "we established a link if at least one of the TRIPOD tests returned a positive coefficient estimate with FDR-adjusted p-values less than 0.01 for at least one trio involving the pair of the TF and the target gene" - this sounds very permissive and it is important to know given this much permissibility, did TRIPOD discover any novel link (hitherto unknown)? Also, for future studies, is this much permissibility recommended?

We believe this permissibility makes sense, because each test in TRIPOD is interrogating a different relationship: Conditional association between TF and target gene, fixing peak association, is different from conditional association between peak association and target gene, fixing TF expression. These are also fundamentally different from interaction effects. Real gene regulatory trios may exhibit any of these relationships, depending on how the mechanics of the regulation. Thus, one major point of this paper is that marginal association does not capture everything. We did find novel links that were validated, for example in Fig. 5c, the light green arrows are newly detected regulatory links by TRIPOD that have support from ChIP-seq data.

Molecular assumptions

15. TRIPOD assumes that the mRNA expression of a TF is a proxy for its protein expression. This is generally not true. Given current data availability, we understand that the authors will need to make this assumption but it is also important to (a) discuss this point and (b) whenever possible, check if protein expression data is available for the TFs discovered in TRIPOD trios (human protein atlas?) and test if this assumption holds for them.

Thanks for this great suggestion. TRIPOD implicitly uses the TF RNA expression as a surrogate for its protein expression. In this revision, we downloaded tissue-specific TF RNA and protein expression levels from the human protein atlas and show that they are highly correlated for both specific TFs (Supplementary Fig. 21a) and across all TFs (Supplementary Fig. 21b). This empirical evidence supports TRIPOD's underlying assumption.

Reviewer #2

The authors here present TRIPOD which is an alternative framework to test the regulatory relations between peaks and gene expression in multi-omic data sets (Figure 2). TRIPOD uses target gene expression, chromatin accessibility of motif containing peaks, and expression of the motif linked transcription factor as 3 variables. Of these 2 are used to in an either-or fashion (always including target gene expression) to establish a regulatory relationship. This relationship is strengthened or weakened using the link with the 3rd variable that wasn't used (level 1). In level 2 analysis the relationship between 2 variables becomes conditional upon the unused relationship and as such more stringent. They show examples of how false positives and false negatives can be corrected for using this method (Figure 2). They benchmark their method to marginal (using 2 variables) analyses and show that in most cases, albeit not all, using a variety of published datasets the analysis is improved (Figures 3-6). In addition, an unexpected link was found between four neurogenesis transcription factors and gliogenic target genes (Figure 5) although little follow up was done.

While others have also used motifs, accessibility and target gene expression, the current method also factors in transcription factor expression which is new (Figure 1). Nevertheless, it's understandable that adding additional variables makes the analysis of regulatory relationships more stringent but generally this comes at a cost of sensitivity. The authors acknowledge this and discuss this handily in the discussion section suggesting it is not the point of the analysis to find everything. However, in using the level 1 analysis I wonder whether much is lost in terms of sensitivity as the analysis works in and either-or fashion adding what is found to the pool of putative regulatory relationships. While no analysis is be perfect, overall, I'm enthusiastic about the approach the authors take and I believe that the strategy the authors use (especially strengthening or weakening relationships using the unused variable) represents an improvement of what is currently being used. Also, the ability to choose the level of analysis is I think an elegant way to deal with sensitivity concerns.

Thank you for your insightful comments about this method. We do concede that TRIPOD is not meant to capture every possible three-way regulatory relationship between TF, their binding site accessibility, and their target genes. This is simply because there are many different types of relationships, of which any model can only be sensitive for a subset. We hope that our paper can clarify and enrich the discussion on this issue, and complement existing models by focusing on trio relationships that might be otherwise missed or falsely called in marginal-based testing.

Perhaps this is what your comment is referring to, and I think it is important to stress: The disagreements between TRIPOD and calls made by marginal-association (and technologies that probe marginal pair-wise relationships, such as ChIP-seq and TF knockout experiments) is not due to differences in sensitivity (or specificity) per se, but due to the models probing fundamentally different models of regulation. The conditional association of Y and Z, *given X*, is fundamentally different from the unconditional association of Y and Z. Thus, our level 1 test is different from existing unconditional association tests. Our level 2 test is, of course, new, since no existing method probes interaction effects. We expect some true relationships to be missed by our method, if they don't involve conditional associations. To clarify this important issue, we have rewritten the following paragraph in the main text:

> *"However, three-way relationships may be complex: When a TF binds to a cis-regulatory region to affect the expression of a gene, it can do so in multiple ways, leading to different patterns in the data. The TF could be directly responsible for opening the chromatin of the enhancer region, facilitating the binding of other TFs that recruit the RNA polymerase. In such cases, expression of the TF is likely to be marginally correlated with the accessibility of the enhancer region, but its correlation with the expression of the target gene may be masked due to confounding of other involved TFs. Alternatively, the TF may not be directly responsible for chromatin remodeling but may bind to already accessible chromatin in recruiting other TFs or the RNA polymerase. In such cases, availability of the TF may not be highly correlated with the accessibility of the enhancer region, but conditioned on the accessibility of the enhancer region, it should be associated with the expression of the target gene. When marginal associations are masked, evidence for binding of the TF at the peak in the regulation of a gene can be inferred from partial associations: (i) with the peak open at a fixed accessibility, whether cells with higher TF expression have higher gene expression; and (ii) with the TF expression fixed at a value above a threshold, whether cells with higher peak accessibility have higher gene expression."*

I have several comments the authors could address.

* The author mentioned in manuscript lines 128:130 "Only peaks within a certain range of the gene's TSS and only interaction between Tfs and peaks containing high-scoring binding motifs for that TFs are considered". It is unclear to me how these threshold/cutoffs are calculated/determined.

> We have carried out our analyses with varying window sizes (50/100/200 kb up/downstream of TSS) and presented the results in Supplementary Fig. 16, using PLAC-seq and ChIP-seq for validation. The other existing methods (e.g., Signac, Cicero, MAESTRO, and ArchR) identify *cis*-regulatory elements and/or compute gene activity scores using the same genomic ranges. Importantly, we show that TRIPOD's performance from the enrichment analysis is invariant to the choice of the window size (Supplementary Fig. 16).

* The author mentioned in manuscript lines 134:135 "To avoid overfitting, we performed out-of-fold prediction and adopted independent training and testing sets". Can the author describe the method more clearly? For example if the author used random forest it should be clear what parameters are used and what the training and test sets are.

> We describe K-fold cross validation in the context of making a prediction (i.e., out-of-fold prediction) in more details in the revised manuscript. Specifically, for the supervised framework (i.e., the LASSO peak and the LASSO TF-peak models), we iteratively hold out $1/K$ of the data, use the remaining data to train the LASSO models, and make predictions on the holdout portion using the estimated coefficients. The results, which are free of overfitting, are shown in Fig. 1b-c.
>
> To further avoid overfitting, we adopt independent training and testing dataset, with results shown in Supplementary Fig. 3. Specifically, we trained the supervised prediction model using the dataset of 10k PBMCs, adopted an independent single-cell multiomic dataset of 3k PBMCs as a testing dataset. We merged the two datasets to match genes, peaks, and TFs, and showed that the peak-TF LASSO model significantly increases the prediction accuracy. We added the details in the methods section of the revised manuscript.

* Lines 141:142 "To mitigate the consequences of sparsity …., we clustered cells to form metacells". Is this nearest neighbor clustering (Seurat)? If so, please describe briefly how it is done. I also assume that this is done for the all-benchmark data sets the same way.

> We added the following details in the Methods section describing how we generate the metacells (renames as cell aggregates following reviewer 1's comment):
>
> "For single-cell data normalization, we use sctransform (Hafemeister and Satija, 2019) and TF-IDF (Stuart et al., 2021) for scRNA-seq and scATAC-seq, respectively, followed by dimension reduction and visualization. To mitigate the effect of ATAC sparsity (Urrutia et al., 2019) and RNA expression stochasticity (Jiang et al., 2017), as a first step, TRIPOD performs cell-wise smoothing by pooling similar cells into "cell aggregates." This, by default, is performed using the WNN method by Seurat V4 (Hao et al., 2021) to jointly reduce dimension and identify cell clusters/states across different modalities. In practice, the cell aggregates can also be inferred using one modality – for example, RNA may better separate the different cell types (La Manno et al., 2021), and in other cases, chromatin

> accessibility may prime cells for differentiation (Ma et al., 2020) – and existing methods such as Seurat (Hafemeister and Satija, 2019) and Signac (Stuart et al., 2021) can be used for domain-specific clustering."
>
> Except for the linkPeaks function by Signac, which takes as input single-cell measurements, we use the same cell-aggregate construction for TRIPOD and all other methods.

* Lines 156:157 are describing how the pipeline deals with LASSO selection behavior. It is known that LASSO selects a maximum of n variables before saturation. How do they deal with selecting the n and also with a group of variables with high pairwise correlation.

> The reviewer is absolutely right that LASSO selects a maximum of n variables if the number of features p is greater than the number of observations n. For TRIPOD, we use the glmnet function to fit the LASSO regression and perform cross-validation to choose the tuning parameter. If multiple variables are highly correlated, LASSO does not perform well in distinguishing which are the "important" ones, as it is geared for prediction. This is why we do not carry out hypothesis testing on LASSO fits, but rather perform the level 1 and level 2 testing of conditional associations via matching. The LASSO prediction shown in Fig. 1 highlights the utility of including both TF expression and *cis*-region accessibility in the model but does not identify significant regulatory relationships, thus mainly acting as a teaser to the actual content of the paper.

* Lines 191:199. As I understand correctly for each metacell-pair a dxp, dyt and dyg is calculated, then metacell-pairs that have low values for each of these measurements, are masked. If this is true, how is this low value determined and to what extend does this method affect the power of the data regarding making it more sparse.

> We do not mask cells that have low values for any of these three values, but only mask cells that have low values for the variable used in matching. We believe that this is a key part of the method: if gene expression and peak region accessibility are correlated only in those cells where the TF is not expressed, then it cannot be a meaningful trio relationship. We agree with the reviewer that the threshold is arbitrary, thus allowing it to be user-specified. By default, the threshold is set to be 0.1 or the 20th percentile of the values for the matched variable, whichever is larger.