nature portfolio

Peer Review File

Cluster replicability in single-cell and single-nucleus atlases of the mouse brain

Corresponding Author: Dr Jesse Gillis

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
French et al. present a comparison of cell clusters identified in two recent large single-cell/single-nucleus transcriptomic atlases of the mouse brain, each comprising >4 million cells grouped into >5,000 clusters. The authors conduct a series of analyses based on marker genes, their MetaNeighbor method, and spatial transcriptomics to identify and characterize cell clusters that are replicable between the two atlases. Key findings include:
1. Both atlases demonstrate similar numbers of cell clusters per dissection region and similar distributions of cluster sizes, suggesting broad agreement on the degree of cellular heterogeneity
2. Marker genes defined by the authors of the original studies often show poor (even below-random) performance in distinguishing clusters from other, transcriptionally similar clusters
3. Using a transcriptome-wide approach (MetaNeighbor), most clusters remain difficult to separate from neighbouring clusters in the other atlas, but thousands of reciprocal best hits can be identified by the AUROC (but not by panels of four or fewer marker genes)
4. Replicable clusters (reciprocal best hits) have a number of distinct technical and biological characteristics: they contain more cells on average; are enriched for particular major cell classes and anatomical regions; and demonstrate higher spatial coherence in independent spatial transcriptomics data (providing a strong independent validation of their methodology). They are distinctly enriched within the cerebellum and depleted in the middle layers of the cerebral cortex
5. Reciprocal best hit clusters can reproduce or inform annotations of reproducible cell types in other datasets, despite variation in experimental procedures or the technical details of the sequencing assay
6. Reciprocal best hit clusters can be identified across the brains of many different species, and cluster replicability is anticorrelated with evolutionary distance

This work addresses an important topic in the field, particularly given the increasing prevalence of extremely large-scale atlas datasets, and the still manual and subjective nature of cell clustering and cluster annotation. Analyses like those described here will be critical if the community is to converge on a common taxonomy of cell types in organs such as the mouse brain. The finding that many clusters are replicable across datasets provides a toehold for further study. At the same time, the finding that a large fraction of clusters reproduce poorly between these studies sounds a note of alarm. All in all, this manuscript is expected to be of high interest to a broad audience. In general, the analyses have been performed rigorously, the claims are well-supported by the data, and the discussion is appropriate. My suggestions are largely intended to improve the clarity of the presentation.

Major points:
1. It would seem that a critical parameter in the authors' analysis is the number of HVGs selected as input to MetaNeighbor. How were the 3,534 HVGs selected? (I could not find a description of this.) How sensitive are the results to the precise number and identity of the HVGs that are provided as input?
2. I did not fully appreciate the implications of the analyses shown in Fig. 4 - how do these relate to the central findings of the study? If the intention is to demonstrate that deeper sampling can identify more reproducible clusters (as implied in the discussion, bottom of p. 15), could this not be demonstrated more directly by counting the number of reciprocal best hits identified in downsampled datasets? If the intention is to further validate the reciprocal best hits, is there a risk of circular reasoning in that the same gene expression data was used to define clusters, identify replicable clusters, and then correlate gene expression across replicable clusters?
3. It would be very helpful to provide a brief overview of MetaNeighbor in the manuscript, so that readers do not need to

consult other papers in order to understand the authors' analysis. (For instance, it is stated on page 12 that "we computed pretrained models" but there is minimal description in the results or methods of what these models are or how they were trained.)
4. In the discussion, the authors propose various biological or technical hypotheses for why certain clusters are less replicable than others, which are plausible. A prominent omission is the possibility that discrete clusters may be an inappropriate way to summarize certain types of transcriptomic variation (cf. Cembrowski and Menon, Trends Neurosci. 2018). A further possibility is that the failure to replicate certain clusters reflects a limitation of computational methods for clustering (i.e., existing methods have identified some spurious clusters). These possibilities would seem worth considering.

Minor points:
5. Many related but distinct analyses of marker genes are described on page 4 and the precise nature of each is not always very clear in the text as written. I suggest the authors consider rewriting this section for clarity.
6. On page 7, it is stated that "Almost all clusters identify another in the other atlas with an AUROC value above 0.95." Do the clusters that do not meet this threshold have any identifiable characteristics? Are they low-quality cells or extremely rare cell types captured in one atlas but not the other?
7. Page 8 describes enrichment of reciprocal best hits among specific classes in the Yao et al. cell type taxonomy. What about depletion? It may be useful to provide a visualization of these results.
8. It was not clear why a logarithmic scale was used for the x-axis in Fig. 5e,c since this distorts the visualization of proportions. If absolute numbers of cells/clusters are difficult to compare across different datasets, it may be preferable to just plot the proportions directly.

(Remarks on code availability)
I did not review the code itself but confirmed it is publicly available and well documented.

Reviewer #2

(Remarks to the Author)
The manuscript, "Cluster replicability in single-cell and single-nucles atlases of the mouse brain," by French et al. assesses the replicability of cell clusters and identified matched cluster pairs across two large mouse brain atlases. The authors assessed cell counts and cell cluster counts across brain regions and found that different experimental technologies capture similar cell heterogeneity. The authors tested marker gene expression specificity in each cell cluster and found that marker genes cannot effectively identify similar clusters but are effective in distinguishing a cluster from the rest of the cells. The authors applied MetaNeighbor to show cluster similarity across datasets by computing the Spearman correlation across highly variable genes. They matched clusters across datasets and analyzed clusters with reciprocal matches in gene expression and spatial location. The authors further applied MetaNeighbor to additional datasets with a wide-range of cell counts and experimental technologies and found reciprocally matched cluster pairs. They also applied MetaNeighbor to brain datasets from 16 species to evaluate replicability and found replicability reduced by evolutionary distance. Overall, the authors used existing methods and much effort to compare brain atlas data and assess the clusters, but could aim for more novel discoveries from these datasets.

Major comments
1. The authors tested marker gene list overlap for every possible pair of cell clusters across datasets. This generated many hypotheses for testing. I would like to see the authors test on only the most similar few pairs of cell clusters (e.g., k-nearest neighbours in average gene expression for each cluster) to see how much the marker genes overlap. The authors also could limit cell cluster comparison to be within the same brain region.
2. The authors pointed out that many multimodal single cell/nucleus cluster integration studies find little to no one-to-one cluster mappings across modalities. Could the authors please explain the reason for focusing on one-to-one cluster mappings? Could several cell states have the same gene expression profile but different profiles in another modality (e.g., chromatin accessibility or methylation)? For example, stem cells can have similar gene expression but are in different chromatin accessibility states so that the cells can differentiate into different cell types. Another example could be that cell cycle creates gene expression changes that follow chromatin accessibility changes to present different cell states within the same cell type.
3. The authors compared two brain atlases. There is inherent variability within one atlas. How robust is the authors' analysis from intrinsic variability? Could the authors consider splitting one brain atlas into two sets of cells and perform the same comparison the author conducted? In an ideal situation, all clusters will have a reciprocal match in gene expression and spatial location. But because of technical variance, this may not be the case. Could the authors comment on what would be the maximum similarity of two datasets, and how does this compare to the results from testing across two different datasets?
4. Could the authors explain the granularity of the cell cluster and subclusters that exist in the brain atlases? How do they compare across datasets? It is currently challenging to quantify the fineness needed for a cell cluster. It would be great to create metrics to assess cell cluster quality.
5. Could the authors please clarify how they matched the cell clusters in Figure 1f (Line 127 to 130)? Are these all possible cluster combinations? It would be great to go into more detail on the reason why there is high correlation in gene expression.
Minor comments
1. Could the authors please include a color scale for the heatmap in Figure 2b?
2. Could the authors explain their reasons for not including analysis with gene markers obtained with a relaxed threshold that can also identify the cluster neighbours? (Line 115 to 116).

(Remarks on code availability)

The github is sufficient.

Reviewer #3

(Remarks to the Author)
French et al. studied the reproducibility of brain cell types at the level of ~5000 fine-grained transcriptomic clusters generated by recent whole mouse brain cell atlases. This seemingly straightforward problem, as the authors showed, is necessary and non-trivial to address. Using a principled statistical framework and an array of rigorous characterizations, the authors provided a few key insights: 1) using the averaged expression of a few marker genes can distinguish a cluster with most other clusters, but this approach usually fails to differentiate a cluster with its closely-related clusters; 2) MetaNeighbor provided a principled and scalable method to evaluate and identify reproducible clusters across datasets ("reciprocal best hits"); 3) although there are examples of remarkable agreements between different atlases, the overall agreement between different studies at the level of fine-grained transcriptomic clusters is only modest; 4) the level of agreement varies depending on brain region (cerebellum vs hypothalamus), gene group (synaptic genes vs others) and phylogeny. These findings are solid, insightful and useful to readers from neurobiology and single-cell omics.

I have no major critiques. Below are some comments that could be further explored at the authors' discretion.

Perhaps it is not surprising that a few marker genes (n=4 seems to be what was used) cannot distinguish ~5000 fine-grained clusters. A typical neuron expresses ~10,000 genes among which a few thousand are highly variable. It is expected that a relatively small set of genes reuse and repurpose in different contexts, rather than being distributed cleanly into unique markers of different cell types. How could this combinatorial and hierarchical nature of gene markers be represented and rigorously tested from a statistical perspective?

The authors showed that MetaNeighbor identified ~2000 reproducible clusters out of the ~5000, that the overall level of agreement between the best vs next clusters is only modest (AUROC ~ 0.6), and that a large block of neuronal clusters have similar AUROCs. I wonder how much of these reflect the limit of the statistical framework proposed here. A few questions could be considered: 1) do we expect closely-related neuronal cell types that are biologically significant (e.g. Economo et al. 2018 Nature) to have meaningful differences in Spearman Correlation across thousands of HVGs derived from a much more broad cell population (including neurons from different regions, glia, and vasculature cells)? 2) can those non-reciprocal best hits be further characterized – some might be one-to-many mapping, some might be different partitions of the same underlying continuous variation (e.g. Xie, Jain, Xu et al. 2025 PNAS), and some might be genuine un-reproducible clusters.

Minor comments:
Fig 1e - are the results for the single-cell or single-nuclei atlas? Text (line 104) suggested both but the figure only has one.
Fig 1f - log CPM - base 2 or base e?
Fig 1h - gave the impression that Liu et al. and the scRNA-seq atlas agree very well in a one-to-one mapped fashion, although the text suggested differently. Could be useful to revise the figure / quantification to reflect the point made in the text, as it is an important point motivating this study.
Line 130 "As expected, some genes are expressed more in whole cells than in nuclei due to the loss of cytoplasmic transcripts." It might be useful to mention that CPM is a relative measure of proportion, and that the reverse is also true that some genes are expressed more in nuclei such as nucleus localized RNAs.

(Remarks on code availability)
Looks well organized. Did not run the code myself.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have carefully and appropriately addressed my comments. The sensitivity analyses of the reciprocal best hits are particularly appreciated since they provide important context for the results (a "replicability ceiling", as the authors note). This is a very nice paper that will be of high interest to the field.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
Thanks to the authors for their detailed response. The authors have addressed most of my concerns. I would like to kindly ask the authors to further clarify several things related to Point 3 in the rebuttal. Please see the list below:
1. Satisfied
2. Satisfied
3. Thanks to the authors for splitting the datasets in half and running MetaNeighbor to match cell clusters. The authors

reported, “Within each atlas, MetaNeighbor correctly and reciprocally paired 80.6% of single-nucleus clusters and 87.9% of single-cell clusters between the halves.” Could the authors please clarify out of which set of clusters the 80.6% and 87.9% of paired clusters come from? And how many clusters total are there for each dataset? How does the split, within-dataset pairing result compare to the number of cross-dataset pairs reported as 2,009? Please clarify the difference between theoretically possible maximum number of pairs and the results from cross-dataset comparison (2,009 pairs). Also, 80% does not seem to be very high for two sets of cells coming from the same dataset. Could the authors please elaborate on this result?
4. Satisfied
5. Satisfied
Minor comments:
1. Satisfied
2. Satisfied

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The authors have done further analysis and discussion to explore the major comments. They also addressed all minor comments. The revised manuscript looks good.

(Remarks on code availability)
The code is publicly accessible and well documented.

> REVIEWER COMMENTS
>
> Reviewer #1 (Remarks to the Author):
>
> French et al. present a comparison of cell clusters identified in two
> recent large single-cell/single-nucleus transcriptomic atlases of the
> mouse brain, each comprising >4 million cells grouped into >5,000
> clusters. The authors conduct a series of analyses based on marker
> genes, their MetaNeighbor method, and spatial transcriptomics to
> identify and characterize cell clusters that are replicable between
> the two atlases. Key findings include:
> 1. Both atlases demonstrate similar numbers of cell clusters per
> dissection region and similar distributions of cluster sizes,
> suggesting broad agreement on the degree of cellular heterogeneity
> 2. Marker genes defined by the authors of the original studies often
> show poor (even below-random) performance in distinguishing clusters
> from other, transcriptionally similar clusters
> 3. Using a transcriptome-wide approach (MetaNeighbor), most clusters
> remain difficult to separate from neighbouring clusters in the other
> atlas, but thousands of reciprocal best hits can be identified by the
> AUROC (but not by panels of four or fewer marker genes)
> 4. Replicable clusters (reciprocal best hits) have a number of
> distinct technical and biological characteristics: they contain more
> cells on average; are enriched for particular major cell classes and
> anatomical regions; and demonstrate higher spatial coherence in
> independent spatial transcriptomics data (providing a strong
> independent validation of their methodology). They are distinctly
> enriched within the cerebellum and depleted in the middle layers of
> the cerebral cortex
> 5. Reciprocal best hit clusters can reproduce or inform annotations of
> reproducible cell types in other datasets, despite variation in
> experimental procedures or the technical details of the sequencing
> assay
> 6. Reciprocal best hit clusters can be identified across the brains of
> many different species, and cluster replicability is anticorrelated
> with evolutionary distance
>
> This work addresses an important topic in the field, particularly
> given the increasing prevalence of extremely large-scale atlas
> datasets, and the still manual and subjective nature of cell
> clustering and cluster annotation. Analyses like those described here
> will be critical if the community is to converge on a common taxonomy
> of cell types in organs such as the mouse brain. The finding that many

> clusters are replicable across datasets provides a toehold for further
> study. At the same time, the finding that a large fraction of clusters
> reproduce poorly between these studies sounds a note of alarm. All in
> all, this manuscript is expected to be of high interest to a broad
> audience. In general, the analyses have been performed rigorously, the
> claims are well-supported by the data, and the discussion is
> appropriate. My suggestions are largely intended to improve the
> clarity of the presentation.
>
> Major points:
> 1. It would seem that a critical parameter in the authors' analysis is
> the number of HVGs selected as input to MetaNeighbor. How were the
> 3,534 HVGs selected? (I could not find a description of this.) How
> sensitive are the results to the precise number and identity of the
> HVGs that are provided as input?

We used the default MetaNeighbor HVG selection method and added its description to the Methods section:
"HVGs were selected with MetaNeighbor's default procedure from the 21,649 genes shared by both atlases: genes were binned into ten deciles by median expression, and within each decile the top 25% by variance were retained. The two atlas-specific highly variable sets were then intersected to yield the final gene set."

We agree that assessing the effect of this parameter is important and have performed additional experiments to evaluate the sensitivity of our results to HVG selection and report the results in the manuscript:

"To test sensitivity to HVG selection, we extracted the same number of HVGs using two alternative methods and also evaluated a random half of the MetaNeighbor HVGs. Using Seurat v3[17], only 38.8% of the genes overlapped with the MetaNeighbor set, yet we recovered a comparable number of reciprocal top hits (1,973), 77.9% of which were identical to the full set. Applying MixHVG[18] (on a random 5% of cells, due to R memory limits) yielded 44.9% gene overlap and 1,899 reciprocal top hits, of which 1,549 matched the full set (77.1%). Randomly removing half of the MetaNeighbor HVGs produced only 10 fewer reciprocal top hits, with 90.6% of them matching the full set of pairs."

These results agree with our coordinated expression results which tested correlation in non-HVGs, and we now mention this agreement in the text:
"The average correlation remained stable across the range of sizes, suggesting that smaller atlases would provide similar levels of coordinated expression. This agrees with our earlier analyses showing that reciprocal top hits are robust to both reduced cell numbers and alternative HVG sets."

> 2. I did not fully appreciate the implications of the analyses shown
> in Fig. 4 - how do these relate to the central findings of the study?
> If the intention is to demonstrate that deeper sampling can identify
> more reproducible clusters (as implied in the discussion, bottom of p.
> 15), could this not be demonstrated more directly by counting the
> number of reciprocal best hits identified in downsampled datasets? If
> the intention is to further validate the reciprocal best hits, is
> there a risk of circular reasoning in that the same gene expression
> data was used to define clusters, identify replicable clusters, and
> then correlate gene expression across replicable clusters?

The coordinated gene expression experiments shown in Figure 4 assessed cluster alignment at the level of genes instead of focusing on specific clusters and their pairs. It also provided information on HVGs importance by testing coordination across the rest of the transcriptome. The exclusion of the HVGs also helped increase the independence of the analyses since it's not the same gene expression data (although the co-expression structure remains). We now mention this directly:
"To avoid biasing the results with the same genes used for pairing, we excluded those 3,534 HVGs, acknowledging that some dependency persists through co-expression."

We agree that we can also assess reciprocal hits in downsampled datasets and have completed that analysis that was also suggested by Reviewer 2. Those results are described alongside the HVG comparisons:

"We next tested how sensitive the reciprocal top hits are to dataset size and HVG selection. First, we randomly split each atlas into two subsets of 2,021,488 cells (corresponding to half of the cells in the smaller single-cell atlas). Within each atlas, MetaNeighbor correctly and reciprocally paired 80.6% of single-nucleus clusters and 87.9% of single-cell clusters between the halves. Across the four comparisons between atlas halves, the number of reciprocal top hits was only slightly lower than in the full analysis (ranging from 1,945 to 1,987), and on average 80.6% of the 2,009 full reciprocal pairs were recovered."

Furthermore, we updated Figure 4e to use a more recent Gene Ontology release, and we now also highlight the children of the Synapse GO terms in addition to the terms themselves.

> 3. It would be very helpful to provide a brief overview of
> MetaNeighbor in the manuscript, so that readers do not need to consult
> other papers in order to understand the authors' analysis. (For
> instance, it is stated on page 12 that "we computed pretrained models"
> but there is minimal description in the results or methods of what

> these models are or how they were trained.)

We agree and have edited parts of the manuscript to give more information on MetaNeighbor. For example, in the context of pretrained models, we added this to the Methods:
“These pretrained models store the HVGs, scaling parameters, and cluster profiles. This allows new datasets to be rapidly projected into the same space and scored for replicability using the atlases as references.”
And in the Results:
“Applying these models to quantify replicability is equivalent to a full MetaNeighbor analysis, but with the HVGs fixed by the atlases.”
Since these pretrained models mirror a normal MetaNeighbor result and represent primarily a technical distinction, we have reduced the number of times we mention ‘pretrained’.

> 4. In the discussion, the authors propose various biological or
> technical hypotheses for why certain clusters are less replicable than
> others, which are plausible. A prominent omission is the possibility
> that discrete clusters may be an inappropriate way to summarize
> certain types of transcriptomic variation (cf. Cembrowski and Menon,
> Trends Neurosci. 2018). A further possibility is that the failure to
> replicate certain clusters reflects a limitation of computational
> methods for clustering (i.e., existing methods have identified some
> spurious clusters). These possibilities would seem worth considering.

We agree that different discretizations of clusters along underlying gradients may explain the lack of agreement between atlases, and we now explicitly address this in the Discussion section:
“In our analyses, many clusters do not match across atlases, possibly due to differences in subcellular and regional sampling, technical variation in sequencing methodologies, or divergent decisions about how to discretize continuous axes of transcriptomic variation into clusters. Previous work has shown that such continuous gradients are present in the brain and can be partitioned differently across studies, potentially contributing to misalignment between atlases (Cembrowski & Menon, 2018; Scala et al., 2021; Xie et al., 2025).”

> Minor points:
> 5. Many related but distinct analyses of marker genes are described on
> page 4 and the precise nature of each is not always very clear in the
> text as written. I suggest the authors consider rewriting this section
> for clarity.

We thank the reviewer for this helpful suggestion. We have rewritten this to more clearly motivate and distinguish the within-atlas marker specificity analyses from the cross-atlas marker overlap analysis. We believe these revisions substantially improve the clarity of this section.

> 6. On page 7, it is stated that “Almost all clusters identify another
> in the other atlas with an AUROC value above 0.95.” Do the clusters
> that do not meet this threshold have any identifiable characteristics?
> Are they low-quality cells or extremely rare cell types captured in
> one atlas but not the other?

We were also curious about these clusters and previously looked into them. We did not find any consistent characteristics that explain their lack of replicability. For example, the two single-nucleus clusters that do not meet the threshold have 1,302 and 6,362 cells, compared to the median of 77. To describe these clusters, we have added the following text to the Results section:

“Exceptions are 29 clusters in the single-cell atlas, which are enriched for oligodendrocyte-lineage clusters (n = 12, pFDR < 0.0001) and intratelencephalic/extratelencephalic glutamatergic neurons (n = 11, pFDR < 0.0001). Only two excitatory single-nucleus neuronal clusters marked by Foxp2, Myzap, and Spsb1 lack a hit with AUROC over 0.95. The median size of these clusters is above the atlas averages, suggesting these are not rare or undersampled populations.”

> 7. Page 8 describes enrichment of reciprocal best hits among specific
> classes in the Yao et al. cell type taxonomy. What about depletion? It
> may be useful to provide a visualization of these results.

We agree and have changed the corresponding Tables S2-3 to report two-sided p-values now, providing readers with tests of both depletion and enrichment. We have edited the text to mention the metaclusters and classes that are depleted for reciprocal best hits:
“We first tested if the reciprocal clusters were enriched or depleted for specific classes from the Yao et al. single-cell based taxonomy. Of the 34 classes, those containing clusters of GABAergic neurons from the medulla, midbrain, and pons; glutamatergic neurons from the medulla and midbrain; and immature GABAergic neurons from the olfactory bulb were enriched for reciprocal hits. In contrast, clusters from the oligodendrocyte-lineage, astrocyte-ependymal, intratelencephalic/extratelencephalic glutamatergic, cerebral nuclei and anterior hypothalamic neurons, and GABAergic neurons from the lateral septal nucleus and hypothalamus were depleted (hypergeometric test, pFDR < 0.05, Supplementary Table 2). Across the meta-cluster classification from the single-nucleus dataset from Langlieb et al. (n=161 with ten or more clusters), astrocytes; and neurons annotated as 'thalamic projection', 'brainstem',

'medullary', 'inhibitory olfactory', 'pontine and medullary', 'tegmental and pontine', and 'midbrain and pontine' were enriched. Depleted meta-clusters were groups of brainstem neuron clusters from the pons and medulla, fibroblasts, and cholinergic neurons of the cranial nerve nuclei (Supplementary Table 3)."

For visualization, we used the spatial data to show the distribution of reciprocal best hits. Specifically, the heatmaps in Figure 3f and 5d show enrichment and depletion at the level of coarse brain regions. More precisely, individual cells mapped to reciprocal pairs are plotted in Figure 3c. Overall, we find these spatial-based visualizations better show the distribution of reciprocal hits, relative to the higher levels of the two taxonomies.

> 8. It was not clear why a logarithmic scale was used for the x-axis in
> Fig. 5e,c since this distorts the visualization of proportions. If
> absolute numbers of cells/clusters are difficult to compare across
> different datasets, it may be preferable to just plot the proportions
> directly.

We agree that the logarithmic scale distorts the x-axis and does not allow accurate estimation of proportions. We have therefore changed these panels to use a linear scale and to plot only the cluster counts, which gives more room to display the full range (14 to 654 clusters in panel c). We also now annotate the proportions directly on the figure to help interpret datasets with small cluster counts. In addition, we report the range of proportions in the Results section.

> Reviewer #1 (Remarks on code availability):
>
> I did not review the code itself but confirmed it is publicly
> available and well documented.

> Reviewer #2 (Remarks to the Author):

>
> The manuscript, "Cluster replicability in single-cell and
> single-nucles atlases of the mouse brain," by French et al. assesses
> the replicability of cell clusters and identified matched cluster
> pairs across two large mouse brain atlases. The authors assessed cell
> counts and cell cluster counts across brain regions and found that
> different experimental technologies capture similar cell
> heterogeneity. The authors tested marker gene expression specificity
> in each cell cluster and found that marker genes cannot effectively
> identify similar clusters but are effective in distinguishing a

> cluster from the rest of the cells. The authors applied MetaNeighbor
> to show cluster similarity across datasets by computing the Spearman
> correlation across highly variable genes. They matched clusters across
> datasets and analyzed clusters with reciprocal matches in gene
> expression and spatial location. The authors further applied
> MetaNeighbor to additional datasets with a wide-range of cell counts
> and experimental technologies and found reciprocally matched cluster
> pairs. They also applied MetaNeighbor to brain datasets from 16
> species to evaluate replicability and found replicability reduced by
> evolutionary distance. Overall, the authors used existing methods and
> much effort to compare brain atlas data and assess the clusters, but
> could aim for more novel discoveries from these datasets.
>
> Major comments
> 1. The authors tested marker gene list overlap for every possible pair
> of cell clusters across datasets. This generated many hypotheses for
> testing. I would like to see the authors test on only the most similar
> few pairs of cell clusters (e.g., k-nearest neighbours in average gene
> expression for each cluster) to see how much the marker genes overlap.

We agree and have compared the author-provided marker lists for all cross-dataset top hit pairs with AUROC > 0.95. This effectively limits the comparison to ~1-2 k-nearest neighbours per cluster and greatly reduces the number of hypotheses relative to testing all possible pairs. We split these overlaps by reciprocal status and now provide the statistics in the Results section:
“Overlap between the author-provided marker lists is limited: only 46% of reciprocal hits share one or more markers (mean overlap of 0.54 genes), and only 25% of non-reciprocal hits have any overlap (0.28 average gene overlap).”

This limited overlap is consistent with our efforts to extract robust markers for the reciprocal pairs (could be described as reciprocal nearest neighbours), as described in the following sentence:
“Beyond the author-provided markers, we used the MetaMarkers framework to select recurrent genes that distinguish both sides of each reciprocal pair; this analysis yielded fewer than 6% of clusters that could be locally discriminated with an AUROC greater than 0.95 using four marker genes (Supplementary Note 1 and Supplementary Figure 1).”

> The authors also could limit cell cluster comparison to be within the
> same brain region.

You are correct, the 23.8 million pairwise cluster comparisons could be reduced. By restricting based on the dissection region as suggested, we can reduce this to 12.2

million by comparing clusters that have cells in at least one shared region. However, that doesn't reduce the number of significant overlaps, as it remains at 30. Applying a more stringent filter that requires the two clusters to have their maximal amount of cells in the same region reduces the multiple-testing burden to 4.8 million comparisons and yields 8 additional significant pairs. We now include this in the Results section text:

"Relaxing the multiple-testing correction to only the 4.8 million cluster pairs that share the same predominant brain region yielded 8 additional significant pairs."

In the following sentence, we also report results from a much more lenient threshold that corrects only for the cluster count in each dataset (~5,000 tests):

"A further, more permissive threshold that corrects only for the cluster count in each dataset (~5,000), identified 1,826 pairs of clusters with overlapping markers."

> 2. The authors pointed out that many multimodal single cell/nucleus
> cluster integration studies find little to no one-to-one cluster
> mappings across modalities. Could the authors please explain the
> reason for focusing on one-to-one cluster mappings?

We mention one-to-one cluster mappings because the atlases are designed to reveal clusters that can serve as stable, named reference cell populations across studies. In contrast, many-to-many mappings make it harder to interpret and communicate correspondences, and complicate analyses. However, we are not restricting our analyses to one-to-one correspondences; rather, we are asking whether the clusters mapped in the three studies overlap at all, irrespective of whether the mapping is one-to-one or not.

> Could several cell
> states have the same gene expression profile but different profiles in
> another modality (e.g., chromatin accessibility or methylation)? For
> example, stem cells can have similar gene expression but are in
> different chromatin accessibility states so that the cells can
> differentiate into different cell types. Another example could be that
> cell cycle creates gene expression changes that follow chromatin
> accessibility changes to present different cell states within the same
> cell type.

We agree that, multiple cell states with similar gene expression profiles could, in principle, be distinguished by chromatin accessibility or DNA methylation assays. However, the resolution of clustering in non-transcriptomic studies is generally coarser than in the RNA-seq data, suggesting that finer-grained states are not being identified.

Some clusters may represent distinct states rather than distinct cell types, but most clusters in these atlases correspond to post-mitotic neurons. Consistent with this, both we and the atlas authors largely refer to cell “clusters” rather than “types” or “states.” For example, in the Yao et al. single-cell atlas, cells collected at different phases of the light–dark cycle primarily commingle within the same clusters. Furthermore, when focusing only on neuronal transcriptomic data, the integration resulted in no one-to-one mappings, indicating that the lack of one-to-one correspondences is not primarily due to modality differences.

> 3. The authors compared two brain atlases. There is inherent
> variability within one atlas. How robust is the authors’ analysis from
> intrinsic variability? Could the authors consider splitting one brain
> atlas into two sets of cells and perform the same comparison the
> author conducted? In an ideal situation, all clusters will have a
> reciprocal match in gene expression and spatial location. But because
> of technical variance, this may not be the case. Could the authors
> comment on what would be the maximum similarity of two datasets, and
> how does this compare to the results from testing across two different
> datasets?

We agree that this is an important point and have now assessed robustness to intrinsic variability by performing six MetaNeighbor comparisons between randomly halved versions of each atlas. We now report this analysis in the Results section:

“We next tested how sensitive the reciprocal top hits are to dataset size and HVG selection. First, we randomly split each atlas into two subsets of 2,021,488 cells (corresponding to half of the cells in the smaller single-cell atlas). Within each atlas, MetaNeighbor correctly and reciprocally paired 80.6% of single-nucleus clusters and 87.9% of single-cell clusters between the halves. Across the four comparisons between atlas halves, the number of reciprocal top hits was only slightly lower than in the full analysis (ranging from 1,945 to 1,987), and on average 80.6% of the 2,009 full reciprocal pairs were recovered. These split-half comparisons establish a replicability ceiling for comparisons between two large, methodologically similar mouse brain atlases.”

> 4. Could the authors explain the granularity of the cell cluster and
> subclusters that exist in the brain atlases? How do they compare
> across datasets? It is currently challenging to quantify the fineness
> needed for a cell cluster.

At the beginning of our Results section we detail the size distribution of the clusters in the brain atlases:

“The cluster sizes in both atlases range from 9 to 264,669 cells in the single-cell dataset and 15 to 552,018 in the single-nucleus dataset. Consequently, a small fraction of the clusters contain most of the cells with a higher concentration of nuclei compared to the cells (Fig. 1c). For example, the ten largest clusters contain 28.9% and 52.4% of the cells in the single-cell and single-nucleus datasets, respectively. Overall, while both atlases provide extensive coverage of the mouse brain, they exhibit differences in cell sampling and cluster distribution, with a significant portion of cells concentrated in a few large clusters.”

We have expanded this paragraph to explain how the two groups clustered the cells:
“Both atlases used recursive graph-based clustering with marker-based stopping rules: the single-cell atlas enforced global separability between all cluster pairs, whereas the single-nucleus atlas retained only clusters with at least three discrete markers relative to other clusters within the same branch of the clustering hierarchy.”

In addition, we have added the below text to the Discussion text in regards to cluster fineness:
“In practice, atlas taxonomies appear most robust when emphasizing larger, replicable cell groupings at the subclass level, annotated by brain region where indicated by spatial distributions.”

> It would be great to create metrics to
> assess cell cluster quality.

We agree that quantitative metrics for cell cluster quality are highly desirable. In this manuscript, we focus primarily on replicability across independent datasets as our central measure of cluster quality, using MetaNeighbor to quantify how reliably clusters are recovered between the single-cell and single-nucleus atlases. We also assess spatial co-localization for the highly replicable clusters.

> 5. Could the authors please clarify how they matched the cell clusters
> in Figure 1f (Line 127 to 130)? Are these all possible cluster
> combinations? It would be great to go into more detail on the reason
> why there is high correlation in gene expression.

Sorry for the confusion, in Fig. 1f the clusters are not matched. Instead, for each gene we compute its average expression across cluster centroids within each atlas. The panel therefore shows the correlation between these atlas-averaged gene expression profiles, effectively comparing overall brain-wide expression between the two studies, which explains the high correlation. We have clarified this in the caption, along with the use of the log base 2 scale as requested by Reviewer 3.

In contrast, in the coordinated expression analysis (Fig. 4), we do use reciprocally matched clusters to correlate genome-wide expression profiles outside of the highly variable genes. This analysis reveals per-gene correlations (average rho=0.357) that are highly specific relative to other genes (95th percentile on average).

> Minor comments
> 1. Could the authors please include a color scale for the heatmap in Figure 2b?

We thank the reviewer for this helpful suggestion, which is particularly important because we use a modified color scale that amplifies the upper range for contrast. The color scale for this panel was previously included but not directly adjacent to the heatmap. We have now enlarged it and moved it closer to make it easier to find (below the bottom-right corner of the heatmap).

> 2. Could the authors explain their reasons for not including analysis
> with gene markers obtained with a relaxed threshold that can also
> identify the cluster neighbours? (Line 115 to 116).

In this section, our goal was to assess how specifically the author-provided markers identify individual clusters and distinguish them from neighbouring clusters. To focus on precise cluster-level identification, we restricted our analyses to the most stringent marker sets. Using the relaxed markers would, by design, collapse neighbouring clusters into shared marker sets, reducing the number of distinct marker sets and blurring the boundaries between closely related clusters, which would lower the marker specificity we measured. We have edited that sentence to make this clear to readers: “However, because our analyses focused on assessing how specifically markers identify individual clusters and distinguish them from all other clusters (including close neighbours), we used only their markers that were obtained without relaxed thresholds.”

> Reviewer #2 (Remarks on code availability):
>
> The github is sufficient.

> Reviewer #3 (Remarks to the Author):
>
> French et al. studied the reproducibility of brain cell types at the
> level of ~5000 fine-grained transcriptomic clusters generated by
> recent whole mouse brain cell atlases. This seemingly straightforward
> problem, as the authors showed, is necessary and non-trivial to
> address. Using a principled statistical framework and an array of
> rigorous characterizations, the authors provided a few key insights:
> 1) using the averaged expression of a few marker genes can distinguish

> a cluster with most other clusters, but this approach usually fails to
> differentiate a cluster with its closely-related clusters; 2)
> MetaNeighbor provided a principled and scalable method to evaluate and
> identify reproducible clusters across datasets (“reciprocal best
> hits”); 3) although there are examples of remarkable agreements
> between different atlases, the overall agreement between different
> studies at the level of fine-grained transcriptomic clusters is only
> modest; 4) the level of agreement varies depending on brain region
> (cerebellum vs hypothalamus), gene group (synaptic genes vs others)
> and phylogeny. These findings are solid, insightful and useful to
> readers from neurobiology and single-cell omics.
>
> I have no major critiques. Below are some comments that could be
> further explored at the authors’ discretion.
>
> Perhaps it is not surprising that a few marker genes (n=4 seems to be
> what was used) cannot distinguish ~5000 fine-grained clusters. A
> typical neuron expresses ~10,000 genes among which a few thousand are
> highly variable. It is expected that a relatively small set of genes
> reuse and repurpose in different contexts, rather than being
> distributed cleanly into unique markers of different cell types. How
> could this combinatorial and hierarchical nature of gene markers be
> represented and rigorously tested from a statistical perspective?

While we haven’t statistically tested complex combinatorial and hierarchical aspects of markers, the Gillis lab has undertaken a statistical study of marker number and metrics in our MetaMarkers work.

We now cite this in the Discussion section:
“Meta-analytic analyses of cortical neuron markers reinforce these limitations, suggesting 50 to 200 markers per cell type (Fischer and Gillis 2021)”

In addition, we used MetaMarkers to identify meta-analytic marker gene lists for the 2,009 reciprocal cluster pairs, as briefly described in the main text and a supplementary note. We’ve added the following sentence to the Results Section to provide marker-count statistics for context:
“The resulting MetaMarkers signatures contain a median of 74 recurrent markers per reciprocal pair (range 15 to 990 genes, with the largest signature for microglia).”

> The authors showed that MetaNeighbor identified ~2000 reproducible
> clusters out of the ~5000, that the overall level of agreement between
> the best vs next clusters is only modest (AUROC ~ 0.6), and that a

> large block of neuronal clusters have similar AUROCs. I wonder how
> much of these reflect the limit of the statistical framework proposed
> here. A few questions could be considered: 1) do we expect
> closely-related neuronal cell types that are biologically significant
> (e.g. Economo et al. 2018 Nature) to have meaningful differences in
> Spearman Correlation across thousands of HVGs derived from a much more
> broad cell population (including neurons from different regions, glia,
> and vasculature cells)?

Other statistical frameworks beyond those we explored may, in principle, better detect differences between closely-related neuron populations. However, in addition to the marker-based analyses and the MetaNeighbor approach we focus on, we note that the three companion studies from the Brain Initiative Cell Census Network 2.0, which employed more complex statistical and integration frameworks, did not appear to achieve clearer or less ambiguous mappings between clusters. Reviewer 1 also raised the issue of HVG selection, and we have performed additional experiments to evaluate the sensitivity of our results to the choice of HVGs and now report these in the manuscript:

“To test sensitivity to HVG selection, we extracted the same number of HVGs using two alternative methods and also evaluated a random half of the MetaNeighbor HVGs. Using Seurat v317, only 38.8% of the genes overlapped with the MetaNeighbor set, yet we recovered a comparable number of reciprocal top hits (1,973), 77.9% of which were identical to the full set. Applying MixHVG18 (on a random 5% of cells, due to R memory limits) yielded 44.9% gene overlap and 1,899 reciprocal top hits, of which 1,549 matched the full set (77.1%). Randomly removing half of the MetaNeighbor HVGs produced only 10 fewer reciprocal top hits, with 90.6% of them matching the full set of pairs.”

Beyond these results, we also tested substantially larger gene sets (up to 7,000 genes), which did not increase the number of reciprocal top hits, and evaluated combined marker-gene sets from each atlas, which yielded only a small (~5%) increase in reciprocal top hits. Together, these analyses suggest that our replicability results are not primarily limited by HVG choice or the MetaNeighbor framework, but instead reflect genuine constraints on reproducibly resolving small cell clusters with current data.

> 2) can those non-reciprocal best hits be
> further characterized

As detailed above in our response to Reviewer 1, we agree and have added a characterization of reciprocal versus non-reciprocal best hits. Specifically, we updated Supplementary Tables 2–3 to report two-sided hypergeometric p-values and added text

describing the classes and meta-clusters that are significantly depleted for reciprocal hits.

> – some might be one-to-many mapping

We spent a significant amount of time and compute resources attempting to merge clusters to increase the number of matches. However, these approaches either reduced coverage or yielded only marginal gains in reciprocal top hits. This is briefly mentioned in the main text and described in more detail in Supplementary Note 2. Early results from a separate effort to integrate the two atlases are consistent with this picture, as the number of clusters in the integrated atlas has increased by over 25%.

> some might
> be different partitions of the same underlying continuous variation
> (e.g. Xie, Jain, Xu et al. 2025 PNAS), and some might be genuine
> un-reproducible clusters.

As mentioned above in response to Reviewer 1, we agree that different discretizations of clusters along underlying gradients may explain the lack of agreement between atlases, and we now explicitly address this in the Discussion section:
“In our analyses, many clusters do not match across atlases, possibly due to differences in subcellular and regional sampling, technical variation in sequencing methodologies, or divergent decisions about how to discretize continuous axes of transcriptomic variation into clusters. Previous work has shown that such continuous gradients are present in the brain and can be partitioned differently across studies, potentially contributing to misalignment between atlases (Cembrowski & Menon, 2018; Scala et al., 2021; Xie et al., 2025).”

> Minor comments:
> Fig 1e - are the results for the single-cell or single-nuclei atlas?

These results are shown for both atlases combined. We considered plotting them separately but did not observe notable differences. This is consistent with the very similar mean AUROCs for the two atlases (0.945 and 0.926 for target versus all other cells, and 0.271 and 0.275 for target versus best off-target). We have edited the figure caption to clarify that the results are shown for both atlases combined.

> Text (line 104) suggested both but the figure only has one.

We have edited this and the second reference to Fig. 1e to clarify that the combined distribution is shown.

> Fig 1f - log CPM - base 2 or base e?

We use base 2 and have updated the axis labels and figure caption to clarify this.

> Fig 1h - gave the impression that Liu et al. and the scRNA-seq atlas
> agree very well in a one-to-one mapped fashion, although the text
> suggested differently. Could be useful to revise the figure /
> quantification to reflect the point made in the text, as it is an
> important point motivating this study.

In Figure 1h, we show the overlap of all clusters between the studies, regardless of whether they are mapped in a one-to-one fashion. Although this is a shallow comparison, the observed overlap is still smaller than expected by chance. We decided not to revise the panel to use a stricter correspondence definition because, in those three studies, mapping to the single-cell atlas was not a central focus.

> Line 130 “As expected, some genes are expressed more in whole cells
> than in nuclei due to the loss of cytoplasmic transcripts.” It might
> be useful to mention that CPM is a relative measure of proportion, and
> that the reverse is also true that some genes are expressed more in
> nuclei such as nucleus localized RNAs.

We thank the reviewer for this helpful suggestion. We have revised the text to clarify:

“In this comparison, most genes fall slightly above the diagonal, indicating higher CPM values in nuclei than in whole cells. This reflects the relative nature of CPM as a measure of proportion: the depletion of cytoplasm-enriched transcripts in the nuclear fraction results in a proportional increase in the remaining genes. Conversely, strongly cytoplasmic transcripts appear as outliers below the diagonal, with higher expression in the single-cell atlas.”

> Reviewer #3 (Remarks on code availability):
>
> Looks well organized. Did not run the code myself.

We thank the reviewers for their helpful comments and are glad that they are satisfied with the revisions to the manuscript. As requested, we have now characterized the clusters that did and did not correctly pair in the within-dataset comparisons. Like Reviewer 2, we also found the interpretation of the within-dataset agreement to be important, and we have provided a point of reference for comparison and expanded this section to fully address the remaining concerns.

> Reviewer #2 (Remarks to the Author):
>
> Thanks to the authors for their detailed response. The authors have
> addressed most of my concerns. I would like to kindly ask the authors to
> further clarify several things related to Point 3 in the rebuttal.
> Please see the list below:
> 1. Satisfied
> 2. Satisfied
> 3. Thanks to the authors for splitting the datasets in half and running
> MetaNeighbor to match cell clusters. The authors reported, “Within each
> atlas, MetaNeighbor correctly and reciprocally paired 80.6% of
> single-nucleus clusters and 87.9% of single-cell clusters between the
> halves.” Could the authors please clarify out of which set of clusters
> the 80.6% and 87.9% of paired clusters come from?

To characterize these clusters, we tested which Metaclusters and Classes were depleted or enriched for the correctly paired cluster halves. These results are now summarized in the text and provided in full in two new supplementary tables:
“Within the broader single-nucleus metaclusters, the correctly paired clusters were enriched in the inhibitory olfactory neuron metacluster, whereas pontine and medullary neuron metaclusters were less likely to match between the halves (Supplementary table 2). At the higher class level in the single-cell dataset, the correct pairs were overrepresented in cortical neuron classes and depleted in the astrocyte-ependymal and classes containing glutamatergic neurons from the pons and medulla (Supplementary table 3).”

> And how many clusters total are there for each dataset?

We have edited this sentence to provide the counts:
“Within each atlas, MetaNeighbor correctly and reciprocally paired 80.6% of single-nucleus clusters (4,054 of 5,030) and 87.9% of single-cell clusters between the halves (4,678 of 5,322).”

> How does the split, within-dataset
> pairing result compare to the number of cross-dataset pairs reported as
> 2,009?

We agree that this is a useful statistic to include and have added it to the Results section:
“The correct within-dataset pairs from the mouse brain atlases were enriched for the 2,009 reciprocal pairs identified in the cross-dataset analyses, recovering 93.9% to 94.6% of them ($p < 0.0001$).”

> Please clarify the difference between theoretically possible
> maximum number of pairs and the results from cross-dataset comparison
> (2,009 pairs).

To clarify this, we now also report the proportion of the 2,009 reciprocal pairs relative to the number of correctly paired clusters in the split-half analyses:
“These pairs establish a practical upper bound for cross-dataset alignment. Relative to these ceilings, the 2,009 reciprocal pair count represents 49.6% of single-nucleus and 42.9% of single-cell clusters.”

> Also, 80% does not seem to be very high for two sets of
> cells coming from the same dataset. Could the authors please elaborate
> on this result?

We agree that this is an important point. Although the split-half comparisons are performed within the same dataset, the 80.6% and 87.9% agreement rates suggest that broader sources of variability limit cluster replicability. As in the cross-dataset analyses, we suspect that this reflects factors such as small clusters and differences in how continuous axes of transcriptomic variation are discretized into clusters. A key point is that these within-dataset agreement rates remain substantially higher than the cross-dataset result, as now highlighted by the comparison above to the 2,009 reciprocal pairs.

To provide a point of reference for these split-half results, we performed the same analysis on a recently released human retina atlas. We did not think this atlas was overclustered based on the number of cells and clusters. Indeed, we find near-complete matching and now report these results in the Results section:
“By comparison, applying the same split-half analysis to a human single-cell retina atlas containing 3.2 million cells showed near-complete recovery, with only one of 123 cell types failing to reciprocally best match between halves.”

We also now refer to this point in the Discussion, in the paragraph suggesting that some over-splitting in the atlases may reflect “technical artifacts or stochastic gene expression differences that do not reflect distinct cellular identities”:
“In practice, atlas taxonomies appear most robust when emphasizing coarser, replicable cell populations. This is supported by our split-half analysis of a retina atlas that

included data from eight studies, in which all but one of the 123 cell groups matched between halves.”

> 4. Satisfied
> 5. Satisfied
> Minor comments:
> 1. Satisfied
> 2. Satisfied