THE EMBO JOURNAL

# Molecular diversity of diencephalic astrocytes reveals adult astrogenesis regulated by Smad4

Stefanie Ohlig, Solene Clavreul, Manja Thorwirth, Tatjana Simon-Ebert, Riccardo Bocchi, Sabine Ulbricht, Nirmal Kannayian, Moritz Rossner, Swetlana Sirko, Pawel Smialowski, Judith Fischer-Sternjak, and Magdalena Götz
**DOI: 10.15252/embj.2020107532**

*Corresponding author(s): Magdalena Götz (magdalena.goetz@helmholtz-muenchen.de) , Swetlana Sirko (swetlana.sirko@med.uni-muenchen.de), Pawel Smialowski (pawel.smialowski@helmholtz-muenchen.de), Judith Fischer-Sternjak (judith.fischer@helmholtz-muenchen.de)*

## Review Timeline:


| | |
|---|---|
| Submission Date: | 16th Dec 20 |
| Editorial Decision: | 2nd Feb 21 |
| Revision Received: | 25th Jun 21 |
| Editorial Decision: | 3rd Aug 21 |
| Revision Received: | 9th Aug 21 |
| Accepted: | 19th Aug 21 |

*Editor: Karin Dumstrei*


## Transaction Report:

(Note: With the exception of the correction of typographical or spelling errors that could be a source of ambiguity, letters and reports are not edited. Depending on transfer agreements, referee reports obtained elsewhere may or may not be included in this compilation. Referee reports are anonymous unless the Referee chooses to sign their reports.)

## 1st Editorial Decision 2nd Feb 2021

Dear Magdalena,

Thank you for submitting your manuscript to The EMBO Journal. I am sorry for the delay in getting back to you with a decision, but I have now received the three referee reports on your manuscript.

As you can see from the comments below, the referees find the analysis interesting and that it provides an important advance in the field. They raise a number of issues that should be fairly straight forward to address. I am therefore happy to invite you to submit a revised manuscript. Let me know if we need to discuss any experiments further - we can do so either via email or video.

When preparing your letter of response to the referees' comments, please bear in mind that this will form part of the Review Process File, and will therefore be available online to the community. For more details on our Transparent Editorial Process, please visit our website:
https://www.embopress.org/page/journal/14602075/authorguide#transparentprocess

I thank you for the opportunity to consider your work for publication. I look forward to your revision.

with best wishes

Karin

Karin Dumstrei, PhD
Senior Editor
The EMBO Journal

Instructions for preparing your revised manuscript:

Please make sure you upload a letter of response to the referees' comments together with the revised manuscript.

Please also check that the title and abstract of the manuscript are brief, yet explicit, even to non-specialists.

When assembling figures, please refer to our figure preparation guideline in order to ensure proper formatting and readability in print as well as on screen:
https://bit.ly/EMBOPressFigurePreparationGuideline

IMPORTANT: When you send the revision we will require
- a point-by-point response to the referees' comments, with a detailed description of the changes made (as a word file).
- a word file of the manuscript text.
- individual production quality figure files (one file per figure)
- a complete author checklist, which you can download from our author guidelines (https://www.embopress.org/page/journal/14602075/authorguide).
- Expanded View files (replacing Supplementary Information)

Please see out instructions to authors
https://www.embopress.org/page/journal/14602075/authorguide#expandedview

Please remember: Digital image enhancement is acceptable practice, as long as it accurately represents the original data and conforms to community standards. If a figure has been subjected to significant electronic manipulation, this must be noted in the figure legend or in the 'Materials and Methods' section. The editors reserve the right to request original versions of figures and the original images that were used to assemble the figure.

Further information is available in our Guide For Authors:
https://www.embopress.org/page/journal/14602075/authorguide

The revision must be submitted online within 90 days; please click on the link below to submit the revision online before 3rd May 2021.

https://emboj.msubmit.net/cgi-bin/main.plex

---

Referee #1:

**Wide-spread and region-specific astrocyte subtypes include proliferative subsets regulated by Smad4**
**Ohlig et al.**

In this paper, Ohlig and colleagues apply a variety of techniques to investigate astrocyte heterogeneity in the mouse diencephalon. Using a combination of single cell transcriptome sequencing and spatial transcriptomics they identify six astrocyte subtypes, five of which appear to show differential positioning in the brain. Interestingly, diencephalic astrocytes show high expression of proliferation-associated genes, suggesting ongoing proliferative ability. This was confirmed by the authors, who demonstrate low levels of ongoing astrocytogenesis using immunostaining, ErdU incorporation and clonal analysis. Using bulk RNA-seq (TRAP) in diencephalic astrocytes (and comparing results to those from non-proliferative cortical astrocytes), the authors identify Smad4 as a potential regulator of proliferation. This was confirmed by gene ablation *in vivo* and in a neurosphere assay. Hence, the authors conclude that mouse diencephalon contains a high degree of astrocyte heterogeneity, including proliferative subtypes regulated by Smad4 signaling.

Overall, this is a high quality paper from one of the leading groups in the astrocyte field. The paper adds substantially to our concept of astrocyte heterogeneity and identifies diencephalon as an area of ongoing (and previously unsuspected) astrocytogenesis. On this conceptual basis, I strongly support publication in EMBO Journal. However, to be acceptable for publication, I feel the authors do need to address the following issues.

**Major issues:**

(i) The spatial transcriptomics data is central to the paper. However, how the integration of single cell RNA-seq data with the spatial (10X Visium) data was performed is poorly described. This is important as the 10X Visium system has limited spatial resolution and is likely to report a complex

molecular signature comprised of multiple cells (and cell types) in the CNS.

Also, is the extrapolation of single cell RNA-seq data from diencephalon to cortex, hippocampus and white matter justifiable? I agree that astrocytes may share some differentially expressed genes between the regions but are they really the same subtype?

(ii) Given the heavy dependence on computational methods in the manuscript, much greater detail should be provided on the choice of parameters used for data analysis (including data integration and clustering), imaging and image processing methods.

For example, the authors claim that their study identifies more genes than the study by Batiuk and colleagues. However, it is not mentioned whether similar methods and thresholds were imposed across the studies. Therefore, please clarify why this claim is justified or rephrase.

(iii) 'Quantification and statistical analysis' in Materials and Methods: Please report all the statistical tests used and why these tests were specifically chosen in each case. In general, it would be good if the statistical methods employed were reported in all appropriate Figure Legends.

(iv) It seems the function(s) of astrocyte subpopulations found by single cell data analysis are assigned manually by curating a few markers. In reality, a full GO analysis on marker genes should be performed and reported.

(v) In my opinion, the authors should provide an easily searchable online tool/database. This is now a common requirment in the field, which would help interested groups to quickly access the data.

**Minor issues:**

(i) The ACSA-2 kit used for astrocyte sorting targets the protein ATP1B2. However, from the tSNE plots this only appears detectable at high levels across a large number of cells in clusters 0 and 5. How do the authors explain this result? In general, I had issues deciding whether genes show zero expression or low levels of expression. Perhaps violin plots would work better than tSNE plots - particularly at the small size used in the Figures? Making the data available online might help in this respect, as investigators can then hunt for their own genes of interest and plot data in a number of ways.....

(ii) The authors claim that cells isolated by MACS were 'mostly' astrocytes. Can this be quantified? Likewise, what does 'virtually identical results' (page 8) mean?

(iii) The authors bring up the concept of 'pan-astrocyte functions' in the 'Introduction'. However, are these really known to be pan-astrocyte functions? Or is it generally just assumed they are pan-astrocytic? It's a subtle distinction in text, but important to frame the significance of the authors' own sequencing data. On page 17, please specify the 'well known astrocyte functions'.

(iv) It is unclear to me whether cells in Cluster 8 are supposed to be considered astrocytes or not. In particular, Page 11 is confusing as it describes (a) Cluster 8 as cells which are not defined as astrocytes and express the NSC marker S100a6 - but also refers to (b) astrocytes with NSC potential? Perhaps, 'astrocyte-like' is an appropriate description for them, which can be used consistently in the text?

(v) Transcript names in mouse are usually given in italics, and this needs to be implemented

throughout the text. References to source studies reporting cell-type specific markers used by authors to identify cell-types should also be added.

(vi) At the end of the first paragraph in the 'Results', the authors refer to Figure 1B. Is this the appropriate Figure, or should it be Figure 1C? Likewise, the text cites 'Panel G' in 'Extended Figure 1' which is not presented. Please check all references to figures in the text.

(vii) The column names in Extended Data Tables 1 and 3 are not clear - they should be more descriptive and not abbreviated. Please also add the total number of genes reported in Extended Data Table 3.

Referee #2:

The manuscript by Ohlig, Clavreul and colleagues focused on characterizing astrocyte heterogeneity in the juvenile diencephalon (DIE). By profiling at single cell resolution (10X Genomics) diencephalon MACS-purified astrocytes from Aldh1l1 BAC-eGFP transgenic line at 8 weeks of age, the authors describe for the first time astrocyte molecular heterogeneity. By analyzing the transcriptional enrichment of the distinct 6 transcriptional clusters resolved, the authors have attempted a regional and functional classification of DIE astrocytes. Overall the data address an interesting and timely topic related to astrocytes diversity, so far almost exclusively explored in few telencephalic regions. On this end, the authors performed a cross-analysis with an already published spatial transcriptomic dataset, and offered insightful observations about DIE astrocytes distribution across the cerebral cortex parenchima and white matter versus the DIE. While intriguing in its concept, in this reviewer's opinion, a deeper computational analysis is required to support both functional and regional heterogeneity. The transcriptional separation also provides evidence for a 'proliferative' signature of DIE astrocytes. The authors performed interesting in vivo birthdating analyisis and lineage-tracing experiments in support of the existence of DIE astrocyte endowed with proliferative features; they also complemented the data with in vitro neurospheres assay and transcriptional bulk analysis. The discovery of a DIE-specific adult astrogeneisis in the intact brain, albeit at very low rate, is quite novel and, in principle, could offer a new outlook at the astrocytes' involvement - for example - in many hypothalamic physiological functions and pathological conditions. It seems to be missing though an experimental paradigm to identify the relevance of this low rate of astrogenesis in the DIE. The manuscript is well written and the figures are clearly represented. The methodologies employed though and the statistical information may benefit of some additional details.

In particular:
1. In the single cell transcriptomic data performed on the MACS-purified astrocytes, it is unclear on what basis the authors defined their k-clustering data and how they determined the diversity degree. High or low degree of molecular heterogeneity (based on number of clusters generated) is difficult to define within the same cellular population. How can the authors discriminate between "state" and "type" in this setting? It is highly intriguing the observation of the most enriched genes in each cluster linked to the potential functional diversity of astrocytes "subtypes", but, in my opinion, it is still highly speculative. First, it is important to bear in mind that transcriptional landscape alone, without epigenomic information or proteomic data, albeit very informative, only partially reflects the cellular activity at any defined moment. In addition, in order to attempt a more robust identification of astrocyte diversity, it would be useful to refer to statistically differential expressed genes in each cluster rather than to cluster markers (only highly enriched and characteristic molecular features of a cluster).

Also, provided the differential levels of expression of canonical astrocyte markers and their hypothesis of low rate steady state proliferation of DIE astrocytes, did the authors perform a pseudotime analysis to trace eventual differentiation trajectories or transitory "states" among the distinct clusters?

2. The cross analysis with spatial transcriptomics is really intriguing and, if confirmed by further investigations, might highlight a new level of regional heterogeneity in CNS astrocytes. In the t-SNE from the published spatial dataset (If available), do any of the clusters resemble the clusters found in from this manuscript's RNA seq? In other words, would those cells also cluster together in the published dataset? To address this question, it might be useful to also computationally extract/isolate astrocytes from the whole spatial analysis and re-cluster them separately to have a higher resolution of the astrocytes. This will be extremely informative on the regional specificity.
Also, what is the consistency of the signature of cluster 0, 1,2,3 in the cerebral cortex and hippocampus dataset? How many of the genes used are found expressed? And in the original clustering, at which ranking position in the marker list would these genes appear (if at all)?
What is the relative percentage of each cluster? Is it correct to define the least abundant are the most region specific? Are those a minority of the astrocyte population in the DIE?
Beyond the spatial data, it could be useful for the authors to also mine additional transcriptional results, with deeper coverage, (i.e. Allen Brain Transcriptional Atlas) to confirm the existence (or lack of thereof in case of the telencephalon) of the populations identified in their own clusters?

3. Would it be possible to identify any combinatorial code of genes (including proliferative ones if needed) to label and quantify those astrocytes in situ and, eventually look for distinct cellular/morphological features?

The proliferative data are supported by multiple genetic strategies and birthdate analysis and overall the argument for proliferative state is convincing. The question remains to me open instead about the number of clones, the rate of division, the survival of the newly born astrocytes and the relevance of such low rate proliferative activity in the intact brain.

4. Does the DIE astrogenesis have any peculiar feature? Is the dynamic any different from the astrogliosis happenings in the injured brain? Could those proliferative astrocytes be the ones that initiate a response in the context of an injury or of a metabolic dysfunction? Why should they be specifically found in the hypothalamus/DIE? Unfortunately, the images in figure 5A are not very clear, and I would recommend to providing orthogonal projections of multiple clones, together with the negative controls (i.e. in different brain regions).

5. Could the authors provide any evidence of the data suggested on page 12 about the survival rate of the adult born DIE astrocytes? Is there any experimental evidence in support to that? If not, I would suggest to simply moving it to the discussion section.

6. The neurospheres data provide puzzling, and, at time slightly disorganized, results, in my opinion. First, can the authors offer additional information (even literature if available) about the culture differences of DIE neurospheres? In absence of EGF and FGF, don't they differentiate at all? Second, could the authors quantitatively assess the pluripotency properties of the DIE-derived NCS in culture? For example, by providing the relative ratio of the three lineages produced out of the number of GLAST-GFP positive cells? Any additional characterization of the different region neurospheres (i. e. diameter size, volume, cellularity) would help exploring the potential differences of the stem cells in the DIE.

7. Lastly, the functional relevance of the proliferative cluster is a critical. Is there any experimental condition, either genetically or metabolically determined, where the functional contribution of those region-restricted astrocytes could be appreciated, or even followed by lineage tracing? High-fat diet animals or genetically obese animals?

Minor points:

1. The first title should be changed, as it is not supported by functional annotation.
2. The detailed information about RNA seq-quality controls and gene selections on page 4 belongs to the method section, rather than the results. I suggest to important information about the clustering results and move elsewhere all the computational controls.
3. When discussing about expression level, please do not refer to it as "qualitative" entities (low and high level of expression), but define the numerical cut-off that has been assigned/used in the analysis for the selection. Although, still relative to the dataset, it is more informative to the reader.
4. I would suggest to implementing the relative figure with the experimental paradigm of EdU injection for clarity.
5. The method section about Edu counts (better if automatized) should be added into the methods, as birthdate counting requires specific analysis of labeling from first and second generation of dividing nuclei.
6. In multiple occasions it is unclear what is considered the experimental 'n'. I strongly encourage the authors to properly clarify in brackets and in the figures on how many animals/trials the experiments were performed and which statistical tests was applied then to the analysis.
7. On page 13, please check for figure/text consistency. Some panels don't seem to add up in the proper order. Panel E in extended figure 5 does not seem to be described anywhere. Please, check the reference to the figures.
8. On page 15 the text states "As we also confirmed the significantly higher levels of Smad4 expression in DIE compared to CTX GM astrocytes isolated by MACS (Figure 8D), we proceeded to explore its functional role."
In the figure the data are not reported as significant. What are the tests applied and the p-value corresponding to the Smad4 expression levels? Also the number of samples should be consistently above thee, where possible and not stated otherwise.

Referee #3:

Ohlig et al. have studied astrocyte heterogeneity in the mouse diencephalon based on scRNAseq analysis as well as the RiboTag technique. They found a segregation between thalamic and hypothalamic astrocytes, and that a small subset of astrocytes retains proliferative and neurosphere-forming potential into adulthood that is in part mediated by Smad4.
One important implication from this study is that individual subsets of astrocytes form the basis of interregional astroglial heterogeneity, rather than broad populations of astrocytes within a given region. Although this hypothesis was not fully explored here, the data nevertheless are novel and interesting, and the study design was rigorous.
I only have a few points:
1. Figure 1 shows that clusters 0 and 5 account for the majority of Atp1b2 expression in diencephalic astrocytes, whereas Atp1b2 expression appears much lower in clusters 1-4. This indicates that cells in clusters 0 and 5 may be overrepresented. Hence, it is debatable whether clusters 1-4 are truly separate clusters and whether the dataset, as well as the absolute number of

4,651 analyzed cells, allows for a full characterization of astrocyte heterogeneity in the diencephalon, especially when considering small clusters.
2. Data presented in Fig. 3 lead the authors to conclude that clusters 1-4 have higher expression levels of genes associated with proliferation, but a closer inspection of Fig. 3 suggests that these differences are marginal at best.
3. What is the reason for the apparent discrepancy between widespread expression of genes associated with proliferation and only <5 % positivity for PCNA or EdU?
4. Are the fluorescently labeled astrocytes in Fig. 5 positive for Ki67 or other markers of proliferation?

Dear Reviewers

Many thanks for your constructive and important comments.
You will see that the manuscript contains a further major revision that none of you had requested – namely control for ependymal cells and accordingly an entire new data set of scRNAseq excluding these.
We were always slightly concerned about possible ependymal cell contamination, as we found many cilia genes in one of the superclusters, but previous work had suggested that ACSA2 would not isolate ependymal cells and we had removed most of the tissue lining third ventricle. However, after we had found a convincing way to label the ependymal cells, namely using CellTracker/Flashtag dye injection into the ventricle, this revealed a substantial presence of ependymal cells in tissue samples dissected for our initial scRNAseq analysis. Using this tracking technique of the entire ventricular system in the brain, it was thus possible to detect and remove the dorsal wall of third ventricle by visual inspection of Flashtag-labelled cells during dissection.
We then repeated the scRNA-seq experiment that now resulted in about 20.000 astrocytes and only minor ependymal cell contamination. The first part of the manuscript is now based on this much improved dataset providing the richest dataset of astrocytes from a single brain region. Following the reviewers` suggestions, we analyzed these by GO term analysis and found some clusters devoted to ion homeostasis, while others were more involved in metabolic regulation or RNA processes. Excitingly, we could still map these different clusters to different positions in the forebrain, but present these data now with more care and details and excluded the respective aspects from title and abstract as suggested by the reviewers. Following the suggestions of the reviewers, we also provide velocity analysis and violin plots of the proliferation genes confirming the wide-spread expression of proliferation genes also in the new dataset. Taken together, this new analysis incorporates all suggestions of reviewers to improve the scRNAseq part on a substantially improved novel data set.

The revisions we have made and our point-by-point responses to the comments are as follows:

Referee #1:

**Wide-spread and region-specific astrocyte subtypes include proliferative subsets regulated by Smad4 Ohlig et al.**

In this paper, Ohlig and colleagues apply a variety of techniques to investigate astrocyte heterogeneity in the mouse diencephalon. Using a combination of single cell transcriptome sequencing and spatial transcriptomics they identify six astrocyte subtypes, five of which appear to show differential positioning in the brain. Interestingly, diencephalic astrocytes show high expression of proliferation-associated genes, suggesting ongoing proliferative ability. This was confirmed by the authors, who demonstrate low levels of ongoing astrocytogenesis using immunostaining, ErdU incorporation and clonal analysis. Using bulk RNA-seq (TRAP) in diencephalic astrocytes (and comparing results to those from non-proliferative cortical astrocytes), the authors identify Smad4 as a potential regulator of proliferation. This was confirmed by gene ablation *in vivo* and in a neurosphere assay. Hence, the authors conclude that mouse diencephalon contains a high degree of astrocyte heterogeneity, including proliferative subtypes regulated by Smad4 signaling.

Overall, this is a high quality paper from one of the leading groups in the astrocyte field. The

paper adds substantially to our concept of astrocyte heterogeneity and identifies diencephalon as an area of ongoing (and previously unsuspected) astrocytogenesis. On this conceptual basis, I strongly support publication in EMBO Journal. However, to be acceptable for publication, I feel the authors do need to address the following issues.

**Major issues:**

(i) The spatial transcriptomics data is central to the paper. However, how the integration of single cell RNA-seq data with the spatial (10X Visium) data was performed is poorly described. This is important as the 10X Visium system has limited spatial resolution and is likely to report a complex molecular signature comprised of multiple cells (and cell types) in the CNS. Also, is the extrapolation of single cell RNA-seq data from diencephalon to cortex, hippocampus and white matter justifiable? I agree that astrocytes may share some differentially expressed genes between the regions but are they really the same subtype?

We have now included a detailed description of the parameters of the scRNA-seq analysis and Visium mapping in the Method section on p. 19 - 20 of the revised manuscript.
We also included the point that the signals in the Visium mapping may include signal from other cells types on p.13 - 14. To have the most reliable data from this analysis we mapped all genes expressed in a cluster, and show that cell types with a known spatial distribution (ependymal cells, blood cells) are positioned as expected (Figure 2, p.5-6).

(ii) Given the heavy dependence on computational methods in the manuscript, much greater detail should be provided on the choice of parameters used for data analysis (including data integration and clustering), imaging and image processing methods.

For example, the authors claim that their study identifies more genes than the study by Batiuk and colleagues. However, it is not mentioned whether similar methods and thresholds were imposed across the studies. Therefore, please clarify why this claim is justified or rephrase.

We have removed this part of the discussion, but included total cell numbers that are now exceeding 20.000 astrocytes in our data set where we find 7 astrocyte clusters/cell states. We also include much more details in the Methods.

(iii) 'Quantification and statistical analysis' in Materials and Methods: Please report all the statistical tests used and why these tests were specifically chosen in each case. In general, it would be good if the statistical methods employed were reported in all appropriate Figure Legends.

We have done so in all Figure legends and also expanded the sections on statistics in the revised Methods part on p.20 - 21.

(iv) It seems the function(s) of astrocyte subpopulations found by single cell data analysis are assigned manually by curating a few markers. In reality, a full GO analysis on marker genes should be performed and reported.

We now present foremost the results of GO enrichment analysis of the scRNAseq as the Source Data for Figure 4, and the top 10 enriched GO terms in clusters 0-4, 9, 10 as a new main Figure 4 as the basis of the different hallmarks of the astrocyte clusters and present these data in the results on p.6-8.

(v) In my opinion, the authors should provide an easily searchable online tool/database. This is now a common requirment in the field, which would help interested groups to quickly access the data.

We are planning to provide an easily searchable tool for the population data. The scRNA data are a bit more tricky to support from our server, but we are searching for options to also provide a searchable tool for this.

**Minor issues:**

(i) The ACSA-2 kit used for astrocyte sorting targets the protein ATP1B2. However, from the tSNE plots this only appears detectable at high levels across a large number of cells in clusters 0 and 5. How do the authors explain this result?

Now we have 2 explanations: First we identified clusters 0-4 as ependymal cells, which may relate to the fact that they have lower levels of ATP1B2, but still sufficient to be isolated by ACSA MACS almost equally well as astrocytes. Second, we know already from Beckervordersandforth et al., 2010 that astroglial cells, neural stem cells and ependymal cells in the adult brain have varying levels of the all known “astroglial” markers. Moreover, many other papers have shown that the levels of e.g. GLAST or other markers vary between astrocytes, so it is not unexpected to have higher and lower levels also for ATP1B2. Indeed, also in the new dataset there is some degree of variation in expression of ATP1B2 across the astrocyte clusters.

In general, I had issues deciding whether genes show zero expression or low levels of expression. Perhaps violin plots would work better than tSNE plots - particularly at the small size used in the Figures?

We now added violin plots for S100a6 in the Figure 3A, for Atp1b2 in Figure EV1E (old dataset) and Appendix S1 (new dataset), respectively.

Making the data available online might help in this respect, as investigators can then hunt for their own genes of interest and plot data in a number of ways.....

The datasets produced in our study are now available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE149115.
In addition, we prepared the list of differentially expressed genes per cluster of scRNA-seq diencephalic astrocytes shown in Source Data for Figure 1B.

(ii) The authors claim that cells isolated by MACS were 'mostly' astrocytes. Can this be quantified? Likewise, what does 'virtually identical results' (page 8) mean?

We avoid such qualitative terms and added the exact number of astrocytes to the text (the number of astrocytes 21503 out of 25761 total cells, p.6 and p.28, respectively).

(iii) The authors bring up the concept of 'pan-astrocyte functions' in the 'Introduction'. However, are these really known to be pan-astrocyte functions? Or is it generally just assumed they are pan-astrocytic?

To our knowledge the evidence for astrocytes taking up glutamate has been obtained in many different regions. Likewise, K-buffering has been seen and shown to be important in many regions, as well as the synaptic function. This is why we call this pan-astrocyte function and we specify this now in the introduction of the revised manuscript.

It's a subtle distinction in text, but important to frame the significance of the authors' own sequencing data. On page 17, please specify the 'well known astrocyte functions'.

We had mentioned them in the introduction as $K^+$ buffering, glutamate up-take, synapse modulation and inserted these examples now also in the discussion.

(iv) It is unclear to me whether cells in Cluster 8 are supposed to be considered astrocytes or not. In particular, Page 11 is confusing as it describes (a) Cluster 8 as cells which are not defined as astrocytes and express the NSC marker S100a6 - but also refers to (b) astrocytes with NSC potential? Perhaps, 'astrocyte-like' is an appropriate description for them, which can be used consistently in the text?

Good point and we now consistently refer to these as astrocyte-like cells.

(v) Transcript names in mouse are usually given in italics, and this needs to be implemented throughout the text. References to source studies reporting cell-type specific markers used by authors to identify cell-types should also be added.

We have now consistently put the transcript names to italics, but are also sometimes referring to protein and added the respective references.

(vi) At the end of the first paragraph in the 'Results', the authors refer to Figure 1B. Is this the appropriate Figure, or should it be Figure 1C? Likewise, the text cites 'Panel G' in 'Extended Figure 1' which is not presented. Please check all references to figures in the text.

We checked and corrected all references to Figures in the manuscript.

(vii) The column names in Extended Data Tables 1 and 3 are not clear - they should be more descriptive and not abbreviated. Please also add the total number of genes reported in Extended Data Table 3.

We indicated the total number of genes on p.5 in the revised manuscript and included the descriptive column names in all source data tables.

Referee #2:

The manuscript by Ohlig, Clavreul and colleagues focused on characterizing astrocyte heterogeneity in the juvenile diencephalon (DIE). By profiling at single cell resolution (10X Genomics) diencephalon MACS-purified astrocytes from Aldh1l1 BAC-eGFP transgenic line at 8 weeks of age, the authors describe for the first time astrocyte molecular heterogeneity. By analyzing the transcriptional enrichment of the distinct 6 transcriptional clusters resolved, the authors have attempted a regional and functional classification of DIE astrocytes. Overall the data address an interesting and timely topic related to astrocytes diversity, so far almost exclusively explored in few telencephalic regions. On this end, the authors performed a cross-analysis with an already published spatial transcriptomic dataset, and offered insightful observations about DIE astrocytes distribution across the cerebral cortex parenchima and white matter versus the DIE. While intriguing in its concept, in this reviewer's opinion, a deeper computational analysis is required to support both functional and regional heterogeneity. The transcriptional separation also provides evidence for a 'proliferative' signature of DIE astrocytes. The authors performed interesting in vivo birthdating analyisis and lineage-tracing experiments in support of the existence of DIE astrocyte endowed with proliferative features; they also complemented the data with in vitro neurospheres assay and transcriptional bulk analysis. The discovery of a DIE-specific adult astrogeneisis in the intact brain, albeit at very low rate, is quite novel and, in principle, could offer a new outlook at the astrocytes' involvement - for example - in many hypothalamic physiological functions and pathological conditions. It seems to be missing though an experimental paradigm to identify the relevance of this low rate of astrogenesis in the DIE. The manuscript is well written and the figures are clearly represented. The methodologies employed though and the statistical information may benefit of some additional details.

In particular:

1. In the single cell transcriptomic data performed on the MACS-purified astrocytes, it is unclear on what basis the authors defined their k-clustering data and how they determined the diversity degree. High or low degree of molecular heterogeneity (based on number of clusters generated) is difficult to define within the same cellular population. How can the authors discriminate between "state" and "type" in this setting?

We now added more details of this in the Methods on p.19-20. We also now avoid the term subtypes throughout the manuscript and clarify in results and discussion that we cannot discriminate between “cell states” versus “subtypes” (e.g. p.13-14). Instead we use the term “gene expression clusters”.

It is highly intriguing the observation of the most enriched genes in each cluster linked to the potential functional diversity of astrocytes "subtypes", but, in my opinion, it is still highly speculative. First, it is important to bear in mind that transcriptional landscape alone, without epigenomic information or proteomic data, albeit very informative, only partially reflects the cellular activity at any defined moment. In addition, in order to attempt a more robust identification of astrocyte diversity, it would be useful to refer to statistically differential expressed genes in each cluster rather than to cluster markers (only highly enriched and characteristic molecular features of a cluster).

We fully agree with the reviewer that the clusters could also indicate different states and now state this clearly in the abstract, results and discussion as mentioned above. We also avoid the term “marker” and rather refer to genes differentially expressed between clusters. Genes for each cluster were calculated using differential gene expression analysis of each cluster versus all remaining data using Wilcox rank sum algorithm. We now specify this more clearly in the methods and results of the revised manuscript.

Also, provided the differential levels of expression of canonical astrocyte markers and their hypothesis of low rate steady state proliferation of DIE astrocytes, did the authors perform a pseudotime analysis to trace eventual differentiation trajectories or transitory "states" among the distinct clusters?

This is an excellent suggestion that we implemented in a new Figure 5C and D. We choose velocity for this analysis as pseudotime requires starting points which one has in development, but not amongst the adult astrocytes. This analysis shows indeed some clusters/cell states with higher dynamism in their velocity plots and also interesting links between clusters/cell states.

2. The cross analysis with spatial transcriptomics is really intriguing and, if confirmed by further investigations, might highlight a new level of regional heterogeneity in CNS astrocytes. In the t-SNE from the published spatial dataset (If available), do any of the clusters resemble the clusters found in from this manuscript's RNA seq? In other words, would those cells also cluster together in the published dataset? To address this question, it might be useful to also computationally extract/isolate astrocytes from the whole spatial analysis and re-cluster them separately to have a higher resolution of the astrocytes. This will be extremely informative on the regional specificity.

Unfortunately, this is not doable as the available spatial data is based on bulk sequencing and hence on cell mixture within confines of spatial resolution (55 micrometer). We can therefore not extract data of different cells.

Also, what is the consistency of the signature of cluster 0, 1,2,3 in the cerebral cortex and hippocampus dataset? How many of the genes used are found expressed? And in the original clustering, at which ranking position in the marker list would these genes appear (if at all)?

We did not select different gene sets from the clusters, but rather use all genes expressed in a cluster for this analysis.

What is the relative percentage of each cluster? Is it correct to define the least abundant are the most region specific? Are those a minority of the astrocyte population in the DIE?

Clusters are numbered according to decreasing size. Exact cell numbers in each cluster are now provided in the Figure legend for 1B-D.
It is an intriguing suggestion that the more region-specific would be the less abundant astrocyte clusters. Indeed, this fits to some, such as cluster 4 (new data), but not cluster 0 with wide-spread mapping while cluster 1 (second most abundant) maps only to the diencephalon. However, altogether only some clusters show region-specific mapping (clusters 1,4 and 9), with most astrocytes in the huge clusters 0, 2 and 3 showing wide-spread mapping.

Beyond the spatial data, it could be useful for the authors to also mine additional transcriptional results, with deeper coverage, (i.e. Allen Brain Transcriptional Atlas) to confirm the existence

(or lack of thereof in case of the telencephalon) of the populations identified in their own clusters?

We have used Visium as it provides the current state-of-the-art resolution and sequencing depth. We are not exactly sure what the reviewer is referring to as deeper coverage of Allen Brain TA. The Visium categories are spatial position which is what we compare to and it is not evident to us to which categories we should compare to from the Allen Brain TA as the spatial resolution is less. However, as we understand the reviewers concern we included a discussion of the spatial mapping in the revised manuscript on p.14.

3. Would it be possible to identify any combinatorial code of genes (including proliferative ones if needed) to label and quantify those astrocytes in situ and, eventually look for distinct cellular/morphological features?

Indeed, these will be future avenues to follow. Here, we focused on functional analysis identifying Smad 4 as a key regulator of the proliferation of diencephalon astrocytes.

The proliferative data are supported by multiple genetic strategies and birthdate analysis and overall the argument for proliferative state is convincing. The question remains to me open instead about the number of clones, the rate of division, the survival of the newly born astrocytes and the relevance of such low rate proliferative activity in the intact brain.

The number of clones is given in the text on p. 9 and following the suggestion of the reviewer, we now provide these data in Source Data for Figure 7, which shows the number of animals and cells analyzed in each animal. The rate of division and survival can only be determined by live imaging, which is very difficult and prone to artefacts in the diencephalon due its localization deep in the murine brain. For live imaging one would have to remove the overlying brain tissue (entire cortex and hippocampus) which may well activate the astrocytes and initiate their proliferation. Therefore at steady state this is not possible and beyond the scope of this manuscript.

4. Does the DIE astrogenesis have any peculiar feature? Is the dynamic any different from the astrogliosis happenings in the injured brain? Could those proliferative astrocytes be the ones that initiate a response in the context of an injury or of a metabolic dysfunction? Why should they be specifically found in the hypothalamus/DIE? Unfortunately, the images in figure 5A are not very clear, and I would recommend to providing orthogonal projections of multiple clones, together with the negative controls (i.e. in different brain regions).

We improved resolution of the overview picture in previous Figure 5A (now Figure 7A) and provide a Z-stack followed by a sequence of individual single plains for each clone example in Figure 7C and D which we think is best suited for showing multiple cells of the same color.
We also show single optical sections of examples where the 2 cells are in the same plane and in cell division to Figure 7E. Again, to follow the dynamics of cell division would require live in vivo imaging with a chronic window as we did after brain injury in the cortex (Bardehle et al., 2013), but unfortunately this is not possible to do without removing the overlying cortex and may hence cause astrocyte activation and proliferation due to this lesion.

5. Could the authors provide any evidence of the data suggested on page 12 about the survival rate of the adult born DIE astrocytes? Is there any experimental evidence in support to that? If not, I would suggest to simply moving it to the discussion section.

The reviewer is right that we did not examine any cell death markers and hence do not know about the survival of the cells, which may anyhow ideally be examined by live imaging, as cell death markers are only detectable for a few hours and the time of death may easily be missed. We therefore followed the suggestion to move this to the discussion.

6. The neurospheres data provide puzzling, and, at time slightly disorganized, results, in my opinion. First, can the authors offer additional information (even literature if available) about the culture differences of DIE neurospheres? In absence of EGF and FGF, don't they differentiate at all?

Indeed, most protocols use EGF and/or FGF2, but it has also been demonstrated that free-floating neurospheres can be obtained when cultured differently. For example, the lab of Verdon Taylor had shown that the Notch ligand Jagged1 can substitute the mitogens EGF and FGF2 and increase the self-renewal and neurogenic potential of NSCs in vitro (Nyfeler et al., 2005). Erickson et al (2008) showed that IGF-1 is necessary for continued passaging of floating neurospheres and that NPC survived for long periods in culture without EGF or FGF-2 when IGF-1 was added to the media. As the Diencephalon neurospheres are new, nobody has published about these before. Indeed, we were surprised to see that they can form neurospheres even in the absence of EGF and FGF2 in contrast to the SVZ cells.
However, given that we detected the presence of ependymal cells in our ACSA2-isolated cell populations, we removed these data from the manuscript and only show the effect of Smad4 on neurosphere formation.

Second, could the authors quantitatively assess the pluripotency properties of the DIE-derived NCS in culture? For example, by providing the relative ratio of the three lineages produced out of the number of GLAST-GFP positive cells?

These data have been removed from the revised manuscript due to the ependymal cell contamination.

Any additional characterization of the different region neurospheres (i. e. diameter size, volume, cellularity) would help exploring the potential differences of the stem cells in the DIE.

In terms of these parameters the diencephalon- and SEZ-derived neurospheres do not differ, but this was never quantified, and we anyhow removed these data from the revised manuscript.

7. Lastly, the functional relevance of the proliferative cluster is a critical. Is there any experimental condition, either genetically or metabolically determined, where the functional contribution of those region-restricted astrocytes could be appreciated, or even followed by lineage tracing? High-fat diet animals or genetically obese animals?

Excellent suggestion and indeed we are just working on the animal license protocol to perform these experiments. Unfortunately, they are not possible within the revision time, but we are very excited to follow this up in future.

Minor points:

1. The first title should be changed, as it is not supported by functional annotation.

We have changed the manuscript title now to: **Molecular diversity of diencephalic astrocytes reveals adult astrogenesis regulated by Smad4.**

2. The detailed information about RNAseq-quality controls and gene selections on page 4 belongs to the method section, rather than the results.

We moved it to the revised methods section on p.19-20.

I suggest to important information about the clustering results and move elsewhere all the computational controls.

We followed this reviewer's suggestion and kept the info about methods for clustering as also asked for by reviewer 1 but moved the remaining info to the methods on p.20.

3. When discussing about expression level, please do not refer to it as "qualitative" entities (low and high level of expression), but define the numerical cut-off that has been assigned/used in the analysis for the selection. Although, still relative to the dataset, it is more informative to the reader.

We have also implemented this suggestion throughout the text.

4. I would suggest to implementing the relative figure with the experimental paradigm of EdU injection for clarity.

We have now included a schematic drawing of the experimental paradigm as Figure 6E.

5. The method section about Edu counts (better if automatized) should be added into the methods, as birthdate counting requires specific analysis of labeling from first and second generation of dividing nuclei.

We did not observe different levels of EdU labelling indicative of consecutive divisions, consistent with the clone size of mostly 2.

6. In multiple occasions it is unclear what is considered the experimental 'n'. I strongly encourage the authors to properly clarify in brackets and in the figures on how many animals/trials the experiments were performed and which statistical tests was applied then to the analysis.

We have carefully scrutinized and edited the manuscript to ensure that the n is inserted everywhere. We also added a list in the methods, besides the Figure legends to list all statistical tests used and indicate for which data which test was used.

7. On page 13, please check for figure/text consistency. Some panels don't seem to add up in the proper order. Panel E in extended figure 5 does not seem to be described anywhere. Please, check the reference to the figures.

We have carefully scrutinized and edited the manuscript to ensure correct reference to the Figure panels in the text.

8. On page 15 the text states "As we also confirmed the significantly higher levels of Smad4 expression in DIE compared to CTX GM astrocytes isolated by MACS (Figure 8D), we proceeded to explore its functional role." In the figure the data are not reported as significant. What are the tests applied and the p-value corresponding to the Smad4 expression levels? Also the number of samples should be consistently above thee, where possible and not stated otherwise.

We now show the values of the biological replicates in EV Figure 7F and also indicate the test used (non-parametric Mann-Whitney test). In addition, we added immunostaining of Smad4 as Figure 9C.

Referee #3:

Ohlig et al. have studied astrocyte heterogeneity in the mouse diencephalon based on scRNAseq analysis as well as the RiboTag technique. They found a segregation between thalamic and hypothalamic astrocytes, and that a small subset of astrocytes retains proliferative and neurosphere-forming potential into adulthood that is in part mediated by Smad4. One important implication from this study is that individual subsets of astrocytes form the basis of interregional astroglial heterogeneity, rather than broad populations of astrocytes within a given region. Although this hypothesis was not fully explored here, the data nevertheless are novel and interesting, and the study design was rigorous.

I only have a few points:
1. Figure 1 shows that clusters 0 and 5 account for the majority of Atp1b2 expression in diencephalic astrocytes, whereas Atp1b2 expression appears much lower in clusters 1-4. This indicates that cells in clusters 0 and 5 may be overrepresented. Hence, it is debatable whether clusters 1-4 are truly separate clusters and whether the dataset, as well as the absolute number of 4,651 analyzed cells, allows for a full characterization of astrocyte heterogeneity in the diencephalon, especially when considering small clusters.

Of course, we do not claim to fully characterize astrocyte heterogeneity in the diencephalon. However, we could now considerably increase our dataset to more than 20.000 astrocytes, thus providing a dataset with the highest number of astrocytes isolated from one region. We have now also added the number of cells in each cluster to the Figure legend 1B-D, which shows that the smallest astrocyte cluster still contains more than 400 cells. We have also added to the text that clusters may of course also represent different cell states in both results and discussion on p. 14.

2. Data presented in Fig. 3 lead the authors to conclude that clusters 1-4 have higher expression levels of genes associated with proliferation, but a closer inspection of Fig. 3 suggests that these differences are marginal at best.

Yes, we fully agree with the reviewer that the differences in proliferation gene expression are not high (except for cluster 8) in this former data set including cluster 1-4 that probably are ependymal cells (see Figure EV3E). As explained in the common section to all reviewers we now succeeded to live label ependymal cells allowing detecting their contribution to the

previous dissection and the removal by dissection for the present dataset. The proliferation index in Figure 5A,B for the new dataset comprising more than 20.000 astrocytes is even more homogenous now and also depicted as violin plot to illustrate this.

3. What is the reason for the apparent discrepancy between widespread expression of genes associated with proliferation and only <5 % positivity for PCNA or EdU?

A long G1 phase, as also shown for oligodendrocyte progenitors proliferating in the adult brain with a division time of 35 to 150 days would explain (Psachoulia et al., 2009), why only few cells are in S-phase at a given time. During a long G1 phase levels of many proliferation associated proteins, such as Ki67 and PCNA fall below detection level. We have now included this information on the bottom of p.8 to clarify the aspects of cell cycle length.

4. Are the fluorescently labeled astrocytes in Fig. 5 positive for Ki67 or other markers of proliferation?

We also tried Ki67 and Mcm5 immunostainings, but found PCNA to be best. As mentioned above, levels are definitely lower than in fast proliferating cells. This may be due to the long cell cycle and long G1 phase, as mentioned on p.8.

## 1st Revision - Editorial Decision 3rd Aug 2021

Dear Magdalena,

Thanks for submitting your revised manuscript to The EMBO Journal. Your study has now been seen by the three referees.

As you can see below, the referees appreciate the introduced changes. Referee #1 has a few remaining comments that I would like to ask you to address with text changes.

When you submit the revised manuscript will you also take care of the following points:

- We don't allow data not shown (pgs 12 + 13) please re-phrase

- We are missing ORCID ID for Smialowski

- The funding information for Collaborative Research Center 870, and SyNergy (EXC 2145 / Projekt-ID 390857198 is missing from the online system.

- Please check that there are figure callouts: for Fig 2D,E,I,K,M-S panels, Fig 4 F,G panels, Fig EV5 panels, EV7 panels and appendix Fig S2 panels

- Fig 2Q panel is missing

- You have 7 EV figures, but can only have 5. The two other figures can be placed in the appendix

- Regarding the source data (list of genes, GO term analysis, RNA seq analysis etc). I think best would be to upload these as .xls files. Please add the legend to the file in a separate tab. The files should be uploaded as dataset files and called Dataset EV1 etc. Let us know if you have any questions regarding this.

- Make sure to make the deposited data publically available (Data Availability Section)

- Our publisher has also done their pre-publication check on your manuscript. When you log into the manuscript submission system you will see the file "Data Edited Manuscript file". Please take a look at the word file and the comments regarding the figure legends and respond to the issues.

- We include a synopsis of the paper (see http://emboj.embopress.org/). Please provide me with a general summary statement and 3-5 bullet points that capture the key findings of the paper.

- We also need a summary figure for the synopsis. The size should be 550 wide by [200-400] high (pixels). You can also use something from the figures if that is easier.

That should be all!

Let me know if you have any further questions

With best wishes

Karin

Karin Dumstrei, PhD
Senior Editor
The EMBO Journal

Guide For Authors: https://www.embopress.org/page/journal/14602075/authorguide

The revision must be submitted online within 90 days; please click on the link below to submit the revision online before 1st Nov 2021.

https://emboj.msubmit.net/cgi-bin/main.plex

---

Referee #1:

**Molecular diversity of diencephalic astrocytes reveals astrogenesis regulated by Smad4. Ohlig et al.**

In this paper, Ohlig and colleagues apply a variety of techniques to investigate astrocyte heterogeneity in the mouse diencephalon. It is a revised version of a manuscript previously submitted to EMBO J as 'Wide-spread and region-specific astrocyte subtypes include proliferative subsets regulated by Smad4'.

The authors use a combination of single cell transcriptome sequencing and spatial transcriptomics to identify multiple astrocyte subtypes/states in the adult mouse diencephalon. Subtypes/states mapped to different positions in the brain; several appeared to be widely distributed and were mapped across the brain to regions including cortex and hippocampus, while others appeared to have a more restricted distribution, centered on diencephalon. Crucially, subtypes/states mapping across the brain showed significant expression of transcripts related to what are commonly perceived to be general astrocyte functions, such as ion transport and ion homeostasis, whereas subtypes mapping in a more restricted fashion appear to be specialized at the transcriptomic level, for example with respect to metabolism. The authors speculate that this may underlie regional matching of astrocytes to local neuronal circuits - an idea which is gaining increasing traction (see Ben Haim and Rowitch, Nat Rev Neurosci, 2017).

Interestingly, diencephalic astrocytes show widespread expression of proliferation-associated genes, suggesting ongoing (and previously unsuspected) proliferative ability. This was confirmed by the authors who demonstrate low levels of ongoing astrocytogenesis using immunostaining, ErdU incorporation and clonal analysis. Using bulk RNA-seq (TRAP) in diencephalic astrocytes (and comparing results to those obtained with non-proliferative cortical astrocytes), the authors identify Smad4 as a potential regulator of proliferation. This was confirmed by gene ablation *in vivo* and in a neurosphere assay.

Hence, the authors conclude that mouse diencephalon contains a high degree of astrocyte heterogeneity, including proliferative subtypes/states regulated by Smad4 signaling. The majority of concerns raised in the initial round of review have been answered satisfactorily. However, the manuscript has been heavily revised to incorporate significant new work during the revision period;

in particular, the single cell sequencing section of the manuscript has been completely revised to exclude data resulting from a previous significant contamination from ependymal cells. This data considerably strengthens this section of the manuscript and the conclusions drawn. However, this also means that there are still issues which require attention/need to be clarified, before the manuscript is acceptable for publication in EMBO Journal.

**Major issues:**
1) In my previous review, I drew attention to the fact that the authors draw strong comparisons to other published single cell sequencing studies. This is also the case in this manuscript version and I am still not convinced it is appropriate, as these studies are technically very different. There is now sufficient published literature showing that 10X Genomics and Smart-seq-based methods are not strictly equivalent (for example, Wang et al., Genomics Proteomics Bioinformatics, 2021). This is before taking into account the fact that previous studies (Batiuk et al., Nat Commun, 2020; Bayraktar et al., Nat Neurosci, 2020) acknowledge that there are likely multiple axes of heterogeneity within astrocyte data, with the total subtype/state number highly dependent on analysis method. Finally, the advantage of sequencing more cells in the current study is offset by the relatively shallow read depth per sample (20K) (Haque et al., Genome Med, 2017). To my mind, these issues mean the 'Discussion' needs to be more nuanced.

In the same manner, the Holt group reported issues with low levels of ACSA-2 staining in other higher-order cell types (oligodendrocytes, mural cells, endothelial cells) in cortex and hippocampus (Batiuk et al., J Biol Chem, 2017; Batiuk et al., Nat Commun, 2020), which might suggest that some level of ependymal cell recovery in diencephalon should not be too surprising.

2) The velocity analysis is, in theory, a nice addition to the manuscript. However, this method works by exploiting differences in the relative abundance of nascent (unspliced) and mature (spliced) mRNA (La Manno et al., Nature, 2018). Along similar lines to Point (1), how the 3' bias inherent to 10X Genomics-based methods affects the obtained results is not discussed (Wang et al., Genomics Proteomics Bioinformatics, 2021). Finally, I am not so convinced of the conclusions: for example, I would not say arrows in cluster 4 are directed towards the main cell cloud. Is it possible to quantify velocity scores?

3) I would encourage the authors to make the data accessible via a searchable webtool as quickly as possible, as I can anticipate a lot of interest from the community in the dataset.

**Minor issues:**
A few small issues remain which, in my opinion, need to be dealt with. Mouse gene names should be routinely italicized. While this has been corrected in the main text, there are still issues with the figures (Figure 1, Figure 3A, Figure 8, Figure 9B, Expanded Figures 1, 3, 5, 7, Appendix Figure S1/S2). In my opinion, the manuscript would also be easier to read (i) if clusters were illustrated in Figure 5 and (ii) tSNE projections of marker genes had cell type of interest indicated on the actual figures.

**Main text:**
1) Page 4, Figure EV1C, D: "*S100b* and *Vim* expression was higher in clusters 1 and 4". I think this should be clusters "1 to 4".

2) Page 7, Figure 5: What criteria were used to select the 9 genes used for calculation of the 'proliferation index'.

**Figures:**

- Figure 1A: The work flow in the diagram will be unclear for people not familiar with the MACS workflow from Batiuk et al., J Biol Chem, 2017. The precise order of steps should be illustrated more clearly or explained more thoroughly in the Figure Legend.

- Figure 2: What does the integration index (scale bar) represent? Please add in the 'Methods' how the segmentation was performed. In addition, a figure showing the degree of overlap between single cell data and spatial data following integration would be a useful 'Expanded View' figure.

- Figure 6: It is unclear if Panel D shows staining against S100 or GFAP (which is also an issue in Figure 9D). Is the S100 antibody supposed to detect all members of the S100 protein family? Why was it used in preference to an antibody specific to S100β, particularly as the Sigma website indicates it was actually raised against S100A10?

- Figure 7: Source Data: Is it correct to refer to 'number of FP+ cell clones', or would 'clonal units' be more appropriate?

- Figure 8:
- Figure 8A: Were the Ribotag mice also expressing GFP as stated in the figure caption? Is 'HA coupling' an appropriate term? Personally, I think something along the lines of 'anti-HA pulldown' is better.
- In terms of showing astrocyte-specific Ribotag expression, Expanded View Figure 7B, C works better than Figure 8B, C (in my opinion) and the authors should consider swapping the figures.
- Figure 8D: are the x-axis counts normalized?
- Figure 8H: how were TFs found/identified?
- Why does the expression normalization change from Figure 8G to Figure 8H?
- Figure 8I: Which methods were used for GO analysis for this figure? Gorilla, Revigo or an overlap of these 2 outputs?

- Figure 9C: Please add a scale bar for the high magnification image.

- Expanded View Figure 1B: Please include the cell type classification after the cluster number, as in Figure 1B. Also, in the figure legend, is it really appropriate to be referring to 'astrocytes' given the high degree of ependymal contamination in the preparation?

- Expanded View Figure 2: Please point out the double positive cells in Panel B with arrows.

**Materials and Methods:**
1) Ribotag experiments: how was read quality checked? Which *Mus musculus* genome was used for mapping - Ensemble....?

2) Single cell and spatial transcriptomics: The authors refer to "subsequent single cell expression analysis" when I assume they mean "clustering analysis on the transcriptomics data"

3) Single cell and spatial transcriptomics: The authors claim they used the "first 11 and 15 dimensions of PCA projection.....to perform clustering". Can the authors please clarify whether they used 11 or 15 PCA, or whether they are referring to analysis of the 'old' and 'new' single cell data, respectively.

4) Quantification and statistical analysis: The authors state that the "normality of data distribution was tested using the Kolmogorov-Smirnov test or Shapiro-Wilk test". Can they please clarify which

test was used on which dataset in the text.

**Acknowledgements:**
Please rephrase the contributions of Jaime Eugenin von Bernhardi and Leda Dimou to the project.

Referee #2:

The authors answered all the raised questions regarding both the computational and the functional analyses. They didn't perform the genetic / metabolic experiment, but this was in fact hardly compatible with the revision timeframe. The work has significantly improved also because the Authors added an important control to exclude ependymal cells from the analysis and have repeated the scRNA-seq experiment excluding this contamination. This made the whole analysis more accurate and clean. I would therefore conclude that they have responded satisfactorily to all the doubts raised.

Referee #3:

The authors have done an excellent job addressing all of my previous points. I have no further comments on this strong paper.

Referee #1:

**Molecular diversity of diencephalic astrocytes reveals astrogenesis regulated by Smad4.**
**Ohlig et al.**

In this paper, Ohlig and colleagues apply a variety of techniques to investigate astrocyte heterogeneity in the mouse diencephalon. It is a revised version of a manuscript previously submitted to EMBO J as 'Wide-spread and region-specific astrocyte subtypes include proliferative subsets regulated by Smad4'.

The authors use a combination of single cell transcriptome sequencing and spatial transcriptomics to identify multiple astrocyte subtypes/states in the adult mouse diencephalon. Subtypes/states mapped to different positions in the brain; several appeared to be widely distributed and were mapped across the brain to regions including cortex and hippocampus, while others appeared to have a more restricted distribution, centered on diencephalon. Crucially, subtypes/states mapping across the brain showed significant expression of transcripts related to what are commonly perceived to be general astrocyte functions, such as ion transport and ion homeostasis, whereas subtypes mapping in a more restricted fashion appear to be specialized at the transcriptomic level, for example with respect to metabolism. The authors speculate that this may underlie regional matching of astrocytes to local neuronal circuits - an idea which is gaining increasing traction (see Ben Haim and Rowitch, Nat Rev Neurosci, 2017).

Interestingly, diencephalic astrocytes show widespread expression of proliferation-associated genes, suggesting ongoing (and previously unsuspected) proliferative ability. This was confirmed by the authors who demonstrate low levels of ongoing astrocytogenesis using immunostaining, ErdU incorporation and clonal analysis. Using bulk RNA-seq (TRAP) in diencephalic astrocytes (and comparing results to those obtained with non-proliferative cortical astrocytes), the authors identify Smad4 as a potential regulator of proliferation. This was confirmed by gene ablation *in vivo* and in a neurosphere assay.

Hence, the authors conclude that mouse diencephalon contains a high degree of astrocyte heterogeneity, including proliferative subtypes/states regulated by Smad4 signaling. The majority of concerns raised in the initial round of review have been answered satisfactorily. However, the manuscript has been heavily revised to incorporate significant new work during the revision period; in particular, the single cell sequencing section of the manuscript has been completely revised to exclude data resulting from a previous significant contamination from ependymal cells. This data considerably strengthens this section of the manuscript and the conclusions drawn. However, this also means that there are still issues which require attention/need to be clarified, before the manuscript is acceptable for publication in EMBO Journal.

**Major issues:**
1) In my previous review, I drew attention to the fact that the authors draw strong comparisons to other published single cell sequencing studies. This is also the case in this manuscript version and I am still not convinced it is appropriate, as these studies are technically very different. There is now sufficient published literature showing that 10X Genomics and Smart-seq-based methods are not strictly equivalent (for example, Wang et al., Genomics Proteomics Bioinformatics, 2021). This is

before taking into account the fact that previous studies (Batiuk et al., Nat Commun, 2020; Bayraktar et al., Nat Neurosci, 2020) acknowledge that there are likely multiple axes of heterogeneity within astrocyte data, with the total subtype/state number highly dependent on analysis method. Finally, the advantage of sequencing more cells in the current study is offset by the relatively shallow read depth per sample (20K) (Haque et al., Genome Med, 2017). To my mind, these issues mean the 'Discussion' needs to be more nuanced.

We politely disagree with the reviewer, as we actually do not mention any direct comparison, but just state, that it is important to collect many astrocytes:

So far in a single adult brain region only astrocytes from the cerebral CTX GM and hippocampus were examined using the same MACS protocol for isolation (Batiuk *et al.*, 2020; Bayraktar *et al.*, 2020) reporting a lower number of clusters of astrocytes with distinct gene expression hallmarks. This shows the importance to collect many astrocytes from one region as done here to achieve sufficient resolution for detecting further differences in gene expression.

However, we have now added:

, even though the sequencing methods are not directly comparable.

In the same manner, the Holt group reported issues with low levels of ACSA-2 staining in other higher-order cell types (oligodendrocytes, mural cells, endothelial cells) in cortex and hippocampus (Batiuk et al., J Biol Chem, 2017; Batiuk et al., Nat Commun, 2020), which might suggest that some level of ependymal cell recovery in diencephalon should not be too surprising.

As ependymal cells were not mentioned in the Batiuk papers, but they were definitely excluded to be sorted by ACSA2 in the Kantzer et al., 2017 manuscript, and we had actually removed most of the ventricular lining, this is why we were surprised. If the reviewer was not surprised, I wonder why he/she did not ask for ependymal cell contamination. We nevertheless revised the statement now to even more clearly refer to the statement in Kantzer et al, that ependymal cells were NOT selected:

“As we had removed most of the ventricular lining by dissection (except the dorsal part indicated in red in Figure 1A), and the first description of ACSA2 selection (Kantzer *et al.*, 2017) had reported that ACSA2 would not select ependymal cells, we were surprised to find high expression of ependymal cell genes in clusters 1-4 (Figure EV3E).”

2) The velocity analysis is, in theory, a nice addition to the manuscript. However, this method works by exploiting differences in the relative abundance of nascent (unspliced) and mature (spliced) mRNA (La Manno et al., Nature, 2018). Along similar lines to Point (1), how the 3' bias inherent to 10X Genomics-based methods affects the obtained results is not discussed (Wang et al., Genomics Proteomics Bioinformatics, 2021). Finally, I am not so convinced of the conclusions: for example, I would not say arrows in cluster 4 are directed towards the main cell cloud. Is it possible to quantify velocity scores?

That is valid consideration and we included sentence alerting the reader to the general 3’ bias of all 10X genomics data. In case of velocity, the genes with low percentage or absent 3’ intro proportion will be affected. To mitigate this issue velocyto (https://pubmed.ncbi.nlm.nih.gov/30089906/)

algorithm calculate velocity per cell taking all detectable genes into account and subsequently average it per cells group as described in MM section.

This information is now added to the results section when describing velocity.

3) I would encourage the authors to make the data accessible via a searchable webtool as quickly as possible, as I can anticipate a lot of interest from the community in the dataset.

As written before in our answers: We will do so for the population analysis as soon as we can and try our best for the single cell analysis.

**Minor issues:**
A few small issues remain which, in my opinion, need to be dealt with. Mouse gene names should be routinely italicized. While this has been corrected in the main text, there are still issues with the figures (Figure 1, Figure 3A, Figure 8, Figure 9B, Expanded Figures 1, 3, 5, 7, Appendix Figure S1/S2). In my opinion, the manuscript would also be easier to read (i) if clusters were illustrated in Figure 5 and (ii) tSNE projections of marker genes had cell type of interest indicated on the actual figures.

We illustrate clusters in Figure 5 now by their numbers and cell type identity.

**Main text:**
1) Page 4, Figure EV1C, D: "*S100b* and *Vim* expression was higher in clusters 1 and 4". I think this should be clusters "1 to 4".

Thank you for noting this. We corrected it now to 1-4.

2) Page 7, Figure 5: What criteria were used to select the 9 genes used for calculation of the 'proliferation index'.

We actually used 2 sets of genes for calculating the proliferation index – a list of 42 taken from other data sets, and a list of 9 which were highest expressed in dividing cells for which we have scRNAseq data. Both sets gave identical results.

**Figures:**
- Figure 1A: The work flow in the diagram will be unclear for people not familiar with the MACS workflow from Batiuk et al., J Biol Chem, 2017. The precise order of steps should be illustrated more clearly or explained more thoroughly in the Figure Legend.

We now added further explanation to the Figure legend.

- Figure 2: What does the integration index (scale bar) represent? Please add in the 'Methods' how the segmentation was performed. In addition, a figure showing the degree of overlap between single cell data and spatial data following integration would be a useful 'Expanded View' figure.

Bars in the figure 2 represent the degree of gene expression similarity between all cells from given cluster and mixture of all cells from given single pixel of Slide-seq data. Figure 2 itself represents

overlap between single cell (10X) and spatial (slide-seq) data in the context of gene expression as measured by 10X genomics technology.

- Figure 6: It is unclear if Panel D shows staining against S100 or GFAP (which is also an issue in Figure 9D). Is the S100 antibody supposed to detect all members of the S100 protein family? Why was it used in preference to an antibody specific to S100β, particularly as the Sigma website indicates it was actually raised against S100A10?

Where we state S100 indeed we used an antibody detecting all S100 proteins. This was used for double stainings with antibodies raised in mice as the S100 antibody was raised in rabbits.

- Figure 7: Source Data: Is it correct to refer to 'number of FP+ cell clones', or would 'clonal units' be more appropriate?

It is correct to refer to as FP+ cell clones.

- Figure 8:
- Figure 8A: Were the Ribotag mice also expressing GFP as stated in the figure caption? Is 'HA coupling' an appropriate term? Personally, I think something along the lines of 'anti-HA pulldown' is better.

We followed the reviewers' suggestion and changed the '@HA- coupling' to 'anti-HA pulldown' in Fig 8A.

- In terms of showing astrocyte-specific Ribotag expression, Expanded View Figure 7B, C works better than Figure 8B, C (in my opinion) and the authors should consider swapping the figures.
We followed the reviewers' suggestion and swapped the Figures accordingly.

- Figure 8D: are the x-axis counts normalized?

Yes, it says normalized counts in the panel and that's what it is.

- Figure 8H: how were TFs found/identified?

To know if a given gene is a transcription factors we were guided by content of AnimalTFDB database (https://pubmed.ncbi.nlm.nih.gov/22080564/)

- Why does the expression normalization change from Figure 8G to Figure 8H?

Panel G is row normalized to highlight similarity between replicates.

- Figure 8I: Which methods were used for GO analysis for this figure? Gorilla, Revigo or an overlap of these 2 outputs?

The figure shows the Revigo data. As stated in the respective Figure legend now, we started the analysis in Gorilla and asked Revigo to use the Gorilla data. Revigo categorizes similar topics using umbrella terms. This results in a shorter list of terms.

- Figure 9C: Please add a scale bar for the high magnification image.

Scale bar has been added.

- Expanded View Figure 1B: Please include the cell type classification after the cluster number, as in Figure 1B. Also, in the figure legend, is it really appropriate to be referring to 'astrocytes' given the high degree of ependymal contamination in the preparation?

We now include the cell type classification also in EV Figure 1B and refer to clusters 1-4 as "astrocytes/ependymal cells/tanycytes".

- Expanded View Figure 2: Please point out the double positive cells in Panel B with arrows.

As virtually all cells are double-positive, we could not indicate them all.

**Materials and Methods:**

1) Ribotag experiments: how was read quality checked? Which *Mus musculus* genome was used for mapping - Ensemble....?

*Mus Musculus Mm10 genome* and was quantified *against* mm10 ensembl release89 annotations.

2) Single cell and spatial transcriptomics: The authors refer to "subsequent single cell expression analysis" when I assume they mean "clustering analysis on the transcriptomics data"

We have now changed the sentence to:

Subsequent single cell expression analysis and clustering was facilitated by Seurat version 2.3.4

3) Single cell and spatial transcriptomics: The authors claim they used the "first 11 and 15 dimensions of PCA projection.....to perform clustering". Can the authors please clarify whether they used 11 or 15 PCA, or whether they are referring to analysis of the 'old' and 'new' single cell data, respectively.

We have now changed the sentence to:

The respective dimensions of PCA projection was used to perform….

4) Quantification and statistical analysis: The authors state that the "normality of data distribution was tested using the Kolmogorov-Smirnov test or Shapiro-Wilk test". Can they please clarify which test was used on which dataset in the text.

We clarified this in the text now.

**Acknowledgements:**
Please rephrase the contributions of Jaime Eugenin von Bernhardi and Leda Dimou to the project.

We have now rephrased this sentence to clarify the contribution of Jaime and Leda.

Referee #2:

The authors answered all the raised questions regarding both the computational and the functional analyses. They didn't perform the genetic / metabolic experiment, but this was in fact hardly compatible with the revision timeframe. The work has significantly improved also because the Authors added an important control to exclude ependymal cells from the analysis and have repeated the scRNA-seq experiment excluding this contamination. This made the whole analysis more accurate and clean. I would therefore conclude that they have responded satisfactorily to all the doubts raised.

Referee #3:

The authors have done an excellent job addressing all of my previous points. I have no further comments on this strong paper.

## 2nd Revision - Editorial Decision 19th Aug 2021

Dear Magdalena,

Thanks for submitting your revised manuscript to The EMBO Journal. I have now had a chance to take a careful look at everything and all looks good!

I am therefore very pleased to accept the manuscript for publication here.

Congratulations on a nice study!

with best wishes

Karin

Karin Dumstrei, PhD
Senior Editor
The EMBO Journal

---

Please note that it is EMBO Journal policy for the transcript of the editorial process (containing referee reports and your response letter) to be published as an online supplement to each paper. If you do NOT want this, you will need to inform the Editorial Office via email immediately. More information is available here:
https://www.embopress.org/page/journal/14602075/authorguide#transparentprocess

Your manuscript will be processed for publication in the journal by EMBO Press. Manuscripts in the PDF and electronic editions of The EMBO Journal will be copy edited, and you will be provided with page proofs prior to publication. Please note that supplementary information is not included in the proofs.

Please note that you will be contacted by Wiley Author Services to complete licensing and payment information. The 'Page Charges Authorization Form' is available here:
https://www.embopress.org/pb-assets/embo-site/tej_apc.pdf

Should you be planning a Press Release on your article, please get in contact with embojournal@wiley.com as early as possible, in order to coordinate publication and release dates.

If you have any questions, please do not hesitate to call or email the Editorial Office. Thank you for your contribution to The EMBO Journal.

** Click here to be directed to your login page: https://emboj.msubmit.net

EMBO PRESS

YOU MUST COMPLETE ALL CELLS WITH A PINK BACKGROUND ↓

PLEASE NOTE THAT THIS CHECKLIST WILL BE PUBLISHED ALONGSIDE YOUR PAPER

| |
|---|
| Corresponding Author Name: Prof. Dr. Magdalena Götz |
| Journal Submitted to: EMBO Journal |
| Manuscript Number: 2020-107532 |

## Reporting Checklist For Life Sciences Articles (Rev. June 2017)

This checklist is used to ensure good reporting standards and to improve the reproducibility of published results. These guidelines are consistent with the Principles and Guidelines for Reporting Preclinical Research issued by the NIH in 2014. Please follow the journal's authorship guidelines in preparing your manuscript.

### A- Figures

### 1. Data

**The data shown in figures should satisfy the following conditions:**

- the data were obtained and processed according to the field's best practice and are presented to reflect the results of the experiments in an accurate and unbiased manner.
- figure panels include only data points, measurements or observations that can be compared to each other in a scientifically meaningful way.
- graphs include clearly labeled error bars for independent experiments and sample sizes. Unless justified, error bars should not be shown for technical replicates.
- if n< 5, the individual data points from each experiment should be plotted and any statistical test employed should be iustified
- Source Data should be included to report the data underlying graphs. Please follow the guidelines set out in the author ship guidelines on Data Presentation.

### 2. Captions

**Each figure caption should contain the following information, for each panel where they are relevant:**

- a specification of the experimental system investigated (eg cell line, species name).
- the assay(s) and method(s) used to carry out the reported observations and measurements
- an explicit mention of the biological and chemical entity(ies) that are being measured.
- an explicit mention of the biological and chemical entity(ies) that are altered/varied/perturbed in a controlled manner.

- the exact sample size (n) for each experimental group/condition, given as a number, not a range;
- a description of the sample collection allowing the reader to understand whether the samples represent technical or biological replicates (including how many animals, litters, cultures, etc.).
- a statement of how many times the experiment shown was independently replicated in the laboratory.
- definitions of statistical methods and measures:
  - common tests, such as t-test (please specify whether paired vs. unpaired), simple $\chi 2$ tests, Wilcoxon and Mann-Whitney tests, can be unambiguously identified by name only, but more complex techniques should be described in the methods section;
  - are tests one-sided or two-sided?
  - are there adjustments for multiple comparisons?
  - exact statistical test results, e.g., P values = x but not P values < x;
  - definition of 'center values' as median or average;
  - definition of error bars as s.d. or s.e.m.

Any descriptions too long for the figure legend should be included in the methods section and/or with the source data.

**In the pink boxes below, please ensure that the answers to the following questions are reported in the manuscript itself. Every question should be answered. If the question is not relevant to your research, please write NA (non applicable). We encourage you to include a specific subsection in the methods section for statistics, reagents, animal models and human subjects.**

## B- Statistics and general methods

Please fill out these boxes ↓ (Do not worry if you cannot see all your text once you press return)

| | |
|---|---|
| 1.a. How was the sample size chosen to ensure adequate power to detect a pre-specified effect size? | The minimum sample size for experiments was determined to allow statistical evaluation of differences between experimental groups with a power higher than 0.86. All statistical data analysis used in this study are described in the 'material and methods' section. |
| 1.b. For animal studies, include a statement about sample size estimate even if no statistical methods were used. | The minimum sample size for experiments was n=3 per genotype/condition. We always used statistical analysis as described in the methods. |
| 2. Describe inclusion/exclusion criteria if samples or animals were excluded from the analysis. Were the criteria pre-established? | No data was excluded in this study. |
| 3. Were any steps taken to minimize the effects of subjective bias when allocating animals/samples to treatment (e.g. randomization procedure)? If yes, please describe. | Not applicable, as the mice/samples we used had different genotypes/regional origin and were allocated according to their genotypes/sources. |
| For animal studies, include a statement about randomization even if no randomization was used. | In this study we used mostly transgenic mice, no pharmacological treatments. We randomized in the sense that we always induced WT or transgenic/mutant animals randomly as they became available. |
| 4.a. Were any steps taken to minimize the effects of subjective bias during group allocation or/and when assessing results (e.g. blinding of the investigator)? If yes please describe. | No specific blinding of the investigators was done, but the genotype of animals or source of cells were typically assessed after the counting´s were completed. |
| 4.b. For animal studies, include a statement about blinding even if no blinding was done | No specific blinding was done, but the animal genotype was typically assessed after the counting´s were completed. |
| 5. For every figure, are statistical tests justified as appropriate? | Each statistical test is justified where used and listed in the methods section. |
| Do the data meet the assumptions of the tests (e.g., normal distribution)? Describe any methods used to assess it. | A description of any assumptions or correction such as tests of normality and adjustment for multiple comparison is justified where used and listed in the methods section. |
| Is there an estimate of variation within each group of data? | Variation within each group of data was taken into consideration by DESEQ2 algorithm calculating p-values for in differential gene expression. |

## USEFUL LINKS FOR COMPLETING THIS FORM

http://www.antibodypedia.com
http://1degreebio.org
http://www.equator-network.org/reporting-guidelines/improving-bioscience-research-repor

http://grants.nih.gov/grants/olaw/olaw.htm
http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/Useofanimals/index.htm
http://ClinicalTrials.gov
http://www.consort-statement.org
http://www.consort-statement.org/checklists/view/32-consort/66-title

http://www.equator-network.org/reporting-guidelines/reporting-recommendations-for-tumo

http://datadryad.org

http://figshare.com

http://www.ncbi.nlm.nih.gov/gap

http://www.ebi.ac.uk/ega

http://biomodels.net/

http://biomodels.net/miriam/
http://jjj.biochem.sun.ac.za
http://oba.od.nih.gov/biosecurity/biosecurity_documents.html
http://www.selectagents.gov/

| | |
|---|---|
| Is the variance similar between the groups that are being statistically compared? | GraphPad Prism (Version 7.03) was used for statistical analysis. The software uses F-test to compare variances while performing t-Test, Mann-Whitney test, Anova, etc. Variance can be different for different genotypes. However, in our study these differences were minimal. |

## C- Reagents

| | |
|---|---|
| 6. To show that antibodies were profiled for use in the system under study (assay and species), provide a citation, catalog number and/or clone number, supplementary information or reference to an antibody validation profile. e.g., Antibodypedia (see link list at top right), 1DegreeBio (see link list at top right). | Data for antibodies and the respective catalog number is provided in the manuscript and methods part. |
| 7. Identify the source of cell lines and report if they were recently authenticated (e.g., by STR profiling) and tested for mycoplasma contamination. | NA |

* for all hyperlinks, please see the table at the top right of the document

## D- Animal Models

| | |
|---|---|
| 8. Report species, strain, gender, age of animals and genetic modification status where applicable. Please detail housing and husbandry conditions and the source of animals. | Data of laboratory animals is provided in the manuscript. All mice were used at young adult age (2 months of age or older) when the experimental treatment began. All sexes were used. They were kept under standard housing conditions with access to water and food ad libitum. We used mostly the GLASTCreERT2 mouse line that we had generated ourselves and described in Mori et al., 2006. This was crossed with the Smad4 floxed/floxed mice (Jackson Laboratories; SMAD4tm2.1CXD/J), Ribotag mice (Jackson Laboratories; B6N.129-Rpl22tm1.1Psam/J), R26R-Confetti reporter mice (Jackson Laboratories, Gt(ROSA)26Sortm1(CAG-Brainbow2.1)Cle/J) and the CAG-eGFP reporter line (FVB.B6-Tg(CAG-cat,-EGFP)1Rbns/KrnzJ). The CAG-eGFP reporter line was generated by Nakamura et al., 2006. Aldh1l1-eGFP mice (Tg(Aldh1l1-EGFP)OFC789Gsat/Mmucd) were rederived from MMRRC |
| 9. For experiments involving live vertebrates, include a statement of compliance with ethical regulations and identify the committee(s) approving the experiments. | Animal handling and experimental procedures were performed in accordance with German and European guidelines and approved by the State of upper Bavaria. |
| 10. We recommend consulting the ARRIVE guidelines (see link list at top right) (PLoS Biol. 8(6), e1000412, 2010) to ensure that other relevant aspects of animal studies are adequately reported. See author guidelines, under 'Reporting Guidelines'. See also: NIH (see link list at top right) and MRC (see link list at top right) recommendations. Please confirm compliance. | We confirm that all animal studies are adequately reported in the manuscript. |

## E- Human Subjects

| | |
|---|---|
| 11. Identify the committee(s) approving the study protocol. | NA |
| 12. Include a statement confirming that informed consent was obtained from all subjects and that the experiments conformed to the principles set out in the WMA Declaration of Helsinki and the Department of Health and Human Services Belmont Report. | NA |
| 13. For publication of patient photos, include a statement confirming that consent to publish was obtained. | NA |
| 14. Report any restrictions on the availability (and/or on the use) of human data or samples. | NA |
| 15. Report the clinical trial registration number (at ClinicalTrials.gov or equivalent), where applicable. | NA |
| 16. For phase II and III randomized controlled trials, please refer to the CONSORT flow diagram (see link list at top right) and submit the CONSORT checklist (see link list at top right) with your submission. See author guidelines, under 'Reporting Guidelines'. Please confirm you have submitted this list. | NA |
| 17. For tumor marker prognostic studies, we recommend that you follow the REMARK reporting guidelines (see link list at top right). See author guidelines, under 'Reporting Guidelines'. Please confirm you have followed these guidelines. | NA |

## F- Data Accessibility

| | |
|---|---|
| 18: Provide a "Data Availability" section at the end of the Materials & Methods, listing the accession codes for data generated in this study and deposited in a public database (e.g. RNA-Seq data: Gene Expression Omnibus GSE39462, Proteomics data: PRIDE PXD000208 etc.) Please refer to our author guidelines for 'Data Deposition'.<br><br>Data deposition in a public repository is mandatory for:<br>a. Protein, DNA and RNA sequences<br>b. Macromolecular structures<br>c. Crystallographic data for small molecules<br>d. Functional genomics data<br>e. Proteomics and molecular interactions | Data availability section for deposited data is provided in the manuscript. |
| 19. Deposition is strongly recommended for any datasets that are central and integral to the study; please consider the journal's data policy. If no structured public repository exists for a given data type, we encourage the provision of datasets in the manuscript as a Supplementary Document (see author guidelines under 'Expanded View' or in unstructured repositories such as Dryad (see link list at top right) or Figshare (see link list at top right). | We will include the following statement in the revised version: Data are available on request to the authors. |
| 20. Access to human clinical and genomic datasets should be provided with as few restrictions as possible while respecting ethical obligations to the patients and relevant medical and legal issues. If practically possible and compatible with the individual consent agreement used in the study, such data should be deposited in one of the major public access-controlled repositories such as dbGAP (see link list at top right) or EGA (see link list at top right). | NA |
| 21. Computational models that are central and integral to a study should be shared without restrictions and provided in a machine-readable form. The relevant accession numbers or links should be provided. When possible, standardized format (SBML, CellML) should be used instead of scripts (e.g. MATLAB). Authors are strongly encouraged to follow the MIRIAM guidelines (see link list at top right) and deposit their model in a public database such as Biomodels (see link list at top right) or JWS Online (see link list at top right). If computer source code is provided with the paper, it should be deposited in a public repository or included in supplementary information. | NA |

## G- Dual use research of concern

| | |
|---|---|
| 22. Could your study fall under dual use research restrictions? Please check biosecurity documents (see link list at top right) and list of select agents and toxins (APHIS/CDC) (see link list at top right). According to our biosecurity guidelines, provide a statement only if it could. | NA |