Author's Response To Reviewer Comments

Reviewer #1: The authors proposed an upgraded Zebrafish Brain Browser atlas, constructed by ANTs SyN registration algorithm, with information from the scans in both the previous atlas construction and the construction of Z-Brain atlas. The registration parameters were optimized separately for both live and fixed tissue scans. Multi-reference channel optimization provided better alignment between Z-Brain and ZBB with better performances in terms of precision and morphology. An additional visualization of the updated atlas was generated for enhanced user experience.

The use of a large number of scans and the application of a more powerful registration algorithm are the main advantages of the upgraded atlas. The results and figures presented promising enhancements compared with the previous version of atlas.

Response: We appreciate the supportive comments.

My specific comments/recommendations mainly concern the registration part:

- In "Methods - zebrafish lines", the sentence in page 4 line 103 "Aside from…" should be moved to registration section.

Response: Done.

- The "Results - optimization of ANTs of live scans" section contained extensive descriptions that should be moved to "Methods - registrations", e.g. the explanation of choosing SyN for registration, the calibration of registration conditions, choice of SyN parameters. The "results" section should present the outcomes of the methods applied, rather than the methods used. The same also applies to "optimization of ANTs for fixed tissue" and "inter-atlas registration".

Response: We moved some text in the Results section to Methods. We also expanded the Methods section in general. However, we think that apart from the main product, the co-alignment of ZBB/Z-Brain, a detailed description of the process for optimization is valuable in the main text, because there are likely to be many other zebrafish datasets that will need to be co-registered. Therefore some of the recommended text (eg chosing SyN for registration) we left as is.

- In page 7 line 205, the MLD was calculated from values given by 3 blinded experts. Was there any inter-rater performance analyses?

Response: Yes - actually we originally selected the 10 points from a larger set of 17 points, on the basis that the experts could reliably locate them on the reference brain. We have added this information to the Methods. However, we are no longer using the 3 blinded experts to identify landmarks in registered brains, because Reviewer 2 asked that we eliminate the cross-correlation metric and focus on MLDs and use 6 brains for optimization. As noted below, it was impractical to have 3 experts locate points in 100's of registered brains. So we now have a single expert locate points in 6 brains before transformation, and simply measure landmark distances after registration from the corresponding reference brain landmarks.

- The whole processes of ANTs registration parameter optimizations, for live scan or fixed tissue or multi-channel combination, can be organized better by using flow charts for presentation. In the current format, it is not conspicuous for readers to follow such processes.

Response: Thank you for this suggestion. We have added three flowcharts (Fig 1a, 3a, 4a) that clarify the procedures for each step.

- In page 10 line 287, were those 167 tERK stained brains part of the 197 scans or new ones? Please clarify.

Response: We have clarified in the text that the 167 tERK brains are a distinct dataset of ours that we used to generate our own average tERK representation for comparison to the Z-Brain average tERK representation (197 tERK brains).

- In page 12 line 359-61, were the MLDs also obtained from 3 experts? Please clarify.

Response: Yes they were, however, as noted above, in switching to new metrics for our analysis, we have removed the manual MLD metric.

- In figure 2, J & K in the label should be j & k.
- In figure 3, Syn in e & f should be SyN. Same for the figure legends.

Response: Thank you, done.

- In figure 4a, please separate M1 & M2 from other values because it is difficult to compare them.

Response: To avoid adding an additional graph with redundant information, we have reduced the number of conditions presented, and made changes to the figure that we think now make it much easier to compare M1 and M2 values.

Reviewer #2: The authors introduce a new method to co-register ZBB and Z-Brain, two existing 6 days post fertilization (dpf) larval zebrafish brain atlases, by using the diffeomorphic algorithm SyN in the ANTs software package. With this method, they provide a quick way to aggregate information from different sources into the same spatial framework creating a comprehensive digital atlas that would provide researchers with a unified resource to gain deeper insights from correlations between neural cell identity, connectivity, gene expression and function within the brain.

Digital brain atlases have been generally restricted to the data produced by the research groups that generated them. In this sense, this work is one of the first efforts to merge two of the main databases for vertebrate larval brain development (6dpf zebrafish brain). From this perspective, the work is novel and of enormous utility to the research community.

Response: Thank you.

The manuscript is sometimes difficult to read and the methods and data are not always properly described. I motivate these points below:

* Through the text it is a bit difficult to follow which datasets (and how many) are used to evaluate registration accuracy and which ones were finally included in the atlas for each of the presented results (registration of live scans, ZBB1.2, registration of fixed scans and Z-Brain/ZBB inter-atlas registration). It could be useful to the reader to expand the "Materials & Methods" and Table 3 to summarize, for each of the mentioned results, which datasets were used as templates, how many brain scans were used for evaluation and how many were mapped to the final resource including the gene patterns they contained, which patterns guided the registration in each case and how many repetitions of each are finally available in the database.

Response: 1. As per the suggestion below and Reviewer 1, we added flowcharts to the figures to make it clear which datasets are used for each experiment and how many volumes were used at each step of the process. 2. We expanded the Methods section. 3. We have also clarified throughout the number of expression patterns that were imported into the database - the total is now 133, including Z-Brain

* From this perspective, it could also help to include a Figure showing, face to face, the average brain representations of all the ZBB patterns that were analyzed in the paper vs. their corresponding Z-Brain counterparts (similarly to Fig.2 in [Marquart et al. 2015]) and a few examples of how they looked before/after registration.

Response: Good idea. We have added a new additional file 4 that shows corresponding ZBB and Z-Brain patterns that were used for registration/verification, and the Z-Brain images after registration.

* A Figure summarizing the final workflow employed could also help the reader: which channels/patterns are finally used to guide the registration, how the rigid and affine steps are used before the elastic transformation, how the alignment is evaluated, etc. (similarly to Fig.3 in [Ronneberger et al. 2012]).

We added flowcharts to the figures to summarize the workflow.

* The "Methods" section describes the elastic registrations performed with ANTs and CMTK but it doesn't mention how the initial rigid and affine transformations are performed.

Response: Table 1 includes the parameters for Rigid and Affine steps for ANTs. We now explicitly state this in the Methods. We amended the text in the Methods to separately clarify the parameters for the Affine (12 dofs, includes rigid transformation) and the Elastic transformations in CMTK.

* The text mentions a "computational measure" on 12 identified landmarks and a "manual

measure" on 10 identified landmarks. However, no details are provided in "Methods" about how those landmarks were identified (automatically? which criteria was given to the 3 blinded reaters?).

Response: The 10 identified landmarks for the blind raters were selected from a larger set in preliminary studies to find landmarks that could be reliably identified. The raters received the same text descriptions as in Additional File 1 and a range of coordinates (within 40 micron cubes, not centered on the location). However, as noted below, we have replaced the MCC metric and manual landmarks with measurements of landmarks before/after registration, so this is not pertinent to the revised manuscript.

In my opinion, some of the evaluation methods employed are not completely aligned to the aims of the study:

* It is a bit confusing why the authors choose to measure (a) mean cross-correlation (MCC) on 50um cubes around 12 landmarks and (b) mean distance to 12 different landmarks. How are the 12 landmarks different to the prior ones? Was it an ad-hoc choice or what was the rationale behind choosing them? Couldn't the same landmarks be used for both measurements? Why are "8 landmarks" and "5-18 landmarks" used for measurement later in the text?

Response: We eliminated these measures, so it is no longer relevant. However, the reason for MCC and MLD originally being on different regions, is because for MCC we used volumes that encompass high-contrast boundaries, whereas for landmarks, we needed specific points that the raters could reliably identify.

* The text mentions that "parameters which yielded the greatest increase in MCC often produce abnormally elongated cells" which seems a strong indication against using MCC as the reference metric for the method. Indeed, [Rohlfing 2012] mentions that "measures such as [...] image similarity [...] do not provide valid evidence for accurate registrations and should thus not be reported or accepted as such". Among these measure of image similarity, [Rohlfing 2012] includes the use of image cross correlation (CC). Restricting CC measurement to a few image regions around landmarks does not solve the problem in my opinion. [Rohlfing 2012] concludes "of the criteria tested in our study, only overlap -(measured as Jaccard index)- of sufficiently local labeled ROIs could distinguish reasonable from poor registrations. One reason for this is that smaller, more localized ROIs approximate point landmarks, and their overlap thus approximates point-based registration error". From this perspective, measuring the mean landmark distance (MLD) in a number of landmarks distributed to cover the image makes much more sense to me: (a) This metric has already been successfully used in prior atlas literature (see [Ronneberger et al. 2012], (b) this metric aligns much better with the goal of "aligning neurons within a cell diameter (~10um)" specified in your Introduction and (c) it is supported by the conclusions in [Rohlfing 2012]. My recommendation is to remove MCC as an evaluation metric and replace it by MLD throughout the paper.

Response: This is an important criticism, thank you. In response, we have completely eliminated MCC as an evaluation metric (except in Fig 3, discussed below) and replaced with MLD. However manually measuring MLD for 100s of registered brains is infeasible. So we adopted a new procedure for MLD that we hope the reviewer will agree is even more rigorous: (1) Identify landmarks in 6 brains before registration and produces a second channel with these positions marked. Does same for the reference. (2) Perform registration of 6 brains to reference, and apply matrix to second channels. (3) Measure MLD using marks in registered second channel, to same points in the reference. Because landmarks are manually annotated in brains before registration, MLDs can then be computationally assessed for the same brains, allowing us to test a large number of registration parameters.

We used this new procedure for optimizing live ANTs registration (Fig. 1), assessing the precision of the new ZBB atlas (Fig. 2) and optimizing Z-Brain registration to ZBB (Fig. 4). The landmarks used in each case are described in new additional figures (1, 3 and 5).

We hope the reviewer agrees that the new metric for point-based registration error, satisfies the concern in Rolhfing et al. We did not include this in the manuscript, but the Reviewer might be interested to learn that there is a strong correlation (R2=0.96) for the MCC for each transformation, with the new MLD measurement.

We were not able to use landmarks for tERK registration (Fig. 3) because, despite Ronneberger 2012, we found it extremely difficult to identify the same specific points along the border of the tectal neuropil in different fish. Yet, this was the very area with extremely poor registration. We therefore retained MCCs, but additionally manually segmented tectal neuropil in order to perform a Jaccard index measurement (as recommended by Rolhfing 2012).

* Additionally, the MCC reported in the results vary considerably from as low as 0.1 to as high as 0.9. There is no indication to the reader about what is the minimum acceptable value to consider a registration appropriate. The dispersion of the results for different brains in Fig.3 also raises questions about the robustness of the method when using different brain scans. In my opinion, these results reflect not only the registration accuracy but also the biological variability between different individuals. From this point of view, it continues to make more sense to me to focus evaluation metrics on MLD where the amount of error that cannot be directly attributed to registration inaccuracies was already quantified (the approximate ~5um differences of blinded raters when labeling landmarks in different datasets).

Response: We agree that MLDs are useful because they provide a clear estimate of registration accuracy. MCCs are easily confounded by salt-and-pepper expression limiting their usefulness in distinguishing between accuracy and biological variability.

* Metrics in the paper should not only be reported as averages but also with their variability. For instance, MLD variability across the 10 different landmarks should be reported (to evaluate the robustness of the method in different parts of the brain). Similarly, MLD variability across different brain scans should be reported (to evaluate the robustness of the method to different datasets). A minimum of 6 brain scans were used in the past to assess such variability [Ronneberger et al. 2012].

Response: Fig. 1 is now based on 6 brain scans. We added variability (for MLD, N=6) for all measures (MLDs, Hausdorff, Elongation index, volume), and the variability specific to each landmark in Additional Files 1 and 3.

* The authors mention the impact the elastic registration has on cell deformation. This is indeed a very relevant point and the qualitative observations performed in the manuscript point in the good direction. However, I feel that these observations are restricted to some anecdotal instances and a more quantitative evaluation may be required to back up claims like "cell morphology remained intact". Similarly to selecting 10 landmarks and measuring their MLD, 10 cells - distributed throughout the brain- could be manually segmented before and after registration to quantitatively measure their deformation (e.g. using the Hausdorff distance [Zanella et al. 2010]). The parameter optimization could then be guided by the objective of achieving an MLD<10um while minimizing cell deformation.

Response: Good idea and we have done exactly this for 107 cells (positions are described in Additional File 2). Fig 1b now shows MLD and Hausdorff distance for the various transformations tested. We do not think that Hausdorff is completely satisfactory, because it is sensitive to the precise alignment of the objects being compared. Thus, although we aligned cell masks before/after registration using their geometric centers and for each mask tested several thousand additional rigid transformations around that point, we were not able to compensate for rotations introduced during registration. We therefore added two supplementary measures of shape (new Figs 1c/d): elongation and volume. Together, these measures quantitatively demonstrate the advantage of the diffeomorphic registration using the optimized parameters.

* Overall, evaluation criteria (MLD, MCC, visually-observed deformation, M1, M2) and number of datasets evaluated (e.g. 6 brains used in Fig.3 vs. 1 brain used in Fig.1a vs. 3 brains employed in Fig.1e-b vs. 167 in Fig.2, etc.) seem to be really heterogeneous throughout the text. It could help to have a consistent unified criteria for the whole paper (e.g. quantitative evaluation of MLD and cell deformation in, say, 2 sets of 3 larvae every time parameters are optimized).

Response: The manuscript is now focussed on MLDs, except for Fig 2 where it was not practical. We retain example images, because they enable the reader to visually compare the effects of optimization. We have used N=6 for Figs 1/3. However for Figs 2/4, this is simply not applicable.

Some of the conclusions are not adequately supported by the data shown,

* ZBB1.2 (with ANTs) is claimed to have "improved registration precision" compared to ZBB (with CMTK). However, the slight improvement reported is not statistically significant. Under these circumstances, it may make more sense to call these results comparable.

Response: After moving from MCCs to MLDs for comparing ZBB and ZBB1.2, the increased precision is significant.

* The claims about cell deformation are based on subjective judgments about registration quality. A more systematic/quantitative approach may be required to generate the supporting evidence.

Response: See above.

Regarding the journal's guidelines on minimum standards of reporting,

* The exact sample size (n) for each experimental group/condition is not clearly reported in the Methods section (see comment above). For instance, for the "fixed registration" and "inter-atlas registration" sections, it is unclear which datasets are used (the 167 brains generated by the group vs. the 197 Z-Brain tERK -which in the intro was reported to contain 899 scans).

* Summary statistics alone are reported sometimes (e.g. aggregate average values) without showing individual data values.

Response: We now provide average, SEM and N throughout.

Minor comments:

* Caption in Fig.2 explains what the arrow points to in (d,e) but not in (f,g), (h,i) and (j,k).

Response: All arrows are now accounted for.

In Fig.3, specify which of the data points in (e,f) corresponds to the dataset shown in (g).

Response: We highlighted the optimal parameter set in f. In the current manuscript we removed (g) because it is anecdotal.