Answers to RC1:

72-73: We completely agree that downhole fractionation is an important effect to be considered in U-Pb dating, which has been shown in many details in the case of zircon, and that it is necessary to be extremely careful with its correction to get the best possible accuracy. However, we observed in the case of carbonate U-Pb dating some differences:

1. Downhole fractionation of NIST 614 and any carbonate material is not comparable, therefore, any attempt to correct a carbonate LA signals using NIST 614 can introduce systematic errors. Even the downhole fractionation trend between different carbonates can be different Fig. RA1
2. Due to the variability of different carbonate minerals and textures, the ablation rate can vary as is shown in the manuscript. Therefore, an attempt to match integration intervals between standard and samples will unfortunately not be perfect as it might be different with respect to downhole fractionation.
3. We agree that if all carbonates showed the same downhole fractionation pattern the approach to get best accuracy is to use exactly the same interval, but due to heterogeneity in ablation rate, within-run variations in initial Pb, U content, and age variations this is hardly possible: a comparison for one piece of JT (10 ablations) is presented comparing total integration with selective and multiple integrations. Fig. RA2
4. If there are high- and low initial Pb zones within a single ablation (which is not uncommon), the total interval will show a large uncertainty with little contribution to a well defined isochron. Two separate integration intervals for high initial Pb and low initial Pb will result in 2 measurement points contributing to a better-defined Isochron, even if the matrix/downhole correction with the total interval from the reference material will add a systematic offset. However, we assume that the variability of initial Pb is randomly distributed from a crater to another and therefore the possible offset is averaged out. Fig. RA2. This might become a problem when the "initial" Pb is mainly at the beginning of the signal due to surface contamination. All precaution should be taken to minimize surface contamination (careful cleaning and pre-ablation).
5. One possible solution would be that only signals where there is minimal variation within the ablation for initial Pb are considered, but then quite a lot of ablations will have to be discarded.
6. We agree on the referee's point that the "adjustments were random enough not to make enough of a difference in the final age". We also agree that it can be possible by selecting only later parts of the signal, the final age will get older respectively younger when only the first parts of the signal are selected. However, the new data of JT, for which we systematically investigated this effect, does not show this behaviour, although data on e.g WC-1 with a clear increasing downhole fractionation trend (Fig RA1) will certainly show this effect. Fig. RA3
7. Figure RA4: shows the downhole trend of WC-1 for different aspect ratios (mean of n>15 signals) both comparing vs time (A) and vs the estimated aspect ratio (B). This figure shows that up to an aspect ratio of 0.2-0.3 no clear downhole trend is observed and only for higher aspect ratios the Pb/U increases significantly.

Therefore, we changed the description accordingly: "The selection of different integration intervals along a single hole ablation can introduce systematic offsets if not randomly distributed due to different amounts of downhole fractionation between RM and sample if there is significant amount of downhole fractionation in either the RM and/or the sample. Best practice is to use as good as possible the same integration intervals with respect to crater shape for both the RM and the sample. As is demonstrated, it is likely that random variability of downhole fractionation, ablation rate, distribution of initial Pb etc. would anyway mitigate the offset potentially introduced. This potentially introduced offset would anyway be diluted in the propagation of the systematic uncertainties, especially since the long term excess variance of secondary RM could precisely result from this.

We improved the manuscript according to some detailed comments.

143: The rather controversially comment "as sometimes done in carbonate U-Pb dating by LA-ICP-MS" has its origin in both personal communications and some unfortunately rather vague descriptions of analytical methods in some publications (like e.g. Nuriel et al., 2017: "either 85 or 110 micron spot size" but no clear statement about whether this is for different sessions or a single one; https://doi.org/10.1130/G38903.1). Therefore, it is of course clearly not suggested in literature to use different crater sizes until now, but this may not prevent users to do so, so we felt like this was important to stress out. To temper this, we adjusted the statement, and can recommend to use different spot sizes and repetition rate as long as the aspect ratio is the same as the reference material.

Figure 4: We adjusted the Y-axis to offset from the ref. value. The associated uncertainties for both the age and the aspect ratio mismatch are large. Especially, the uncertainty of the aspect ratio mismatch relative to the RM is difficult to quantify. Indeed, these values result from a combination of estimations based on the number of pulses and an average ablation rate for this sample, itself based on measurements made for some (but not all) craters in some (but not all) sessions. We think that measuring the depth of each individual crater as precisely as possible would be a considerable amount of work considering the small influence that it would have on the final results. Therefore, we do not give an uncertainty on the x value and also refrain from calculating an uncertainty on the slope. Based on the new slopes, we think they are equal within uncertainty.
Figure 4 is also meant to show the general effect observed and described in this manuscript that with changes of the crater geometry you likely get significant age offsets. However, the ablation rate of the carbonates is extremely variable (as shown in Figure 5) and can vary both from spot to spot of the sample sample, as we occasionally observed. The figure qualitatively shows that the described effect is present, and that with different aspect ratios an offset is introduced.

What the referee 1 suggests is in the end a perfect matrix matching between RM and sample, which considering all the sources of uncertainty above, is in our opinion not possible. What we can suggest is therefore a reasonable approach to get as accurate results as possible for many interesting applications with the presented approach.

To be able to adjust the crater size is an advantage that makes LA-ICP-MS the method of choice for many applications that we would not sacrifice this versatility, but obviously care has to be taken. If the grain size of a sample is highly variable we like to apply different crater sizes to get trace element content for low concertation with larger spot size and good reproducibility for higher trace elements with many replicates of small craters. In the end there have to be some compromises and we think that in the manuscript we show a way to use different sample sizes without sacrificing the accuracy too much by adjusting as good as possible the crater geometry. We think the dependence of the uncertainty on the signal intensity is well shown in literature and with larger craters a higher ablation rate more sensitivity gives smaller uncertainty.

In the end, we are confident that using the excess uncertainty as recommended in this manuscript would anyway cover the uncertainty associated with the differences between RM and most samples, including differences in downhole fractionation, ablation rate, integration intervals, etc. – as long as the aspect ratio of the craters is kept similar and the whole analytical protocol is followed as good as possible.

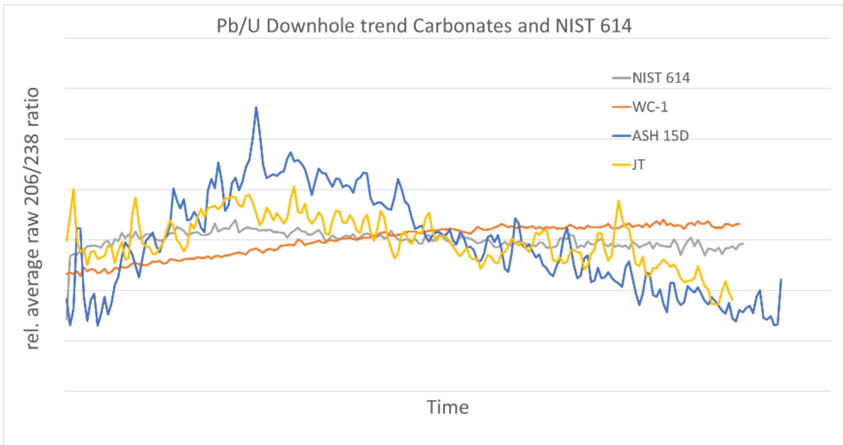

Figure RA1: Relative to the mean raw 206/238 ratio) downhole fractionation trend for 3 different Carbonates and NIST 614 showing no consistent trend, making a downhole fractionation correction of unknown samples very difficult.

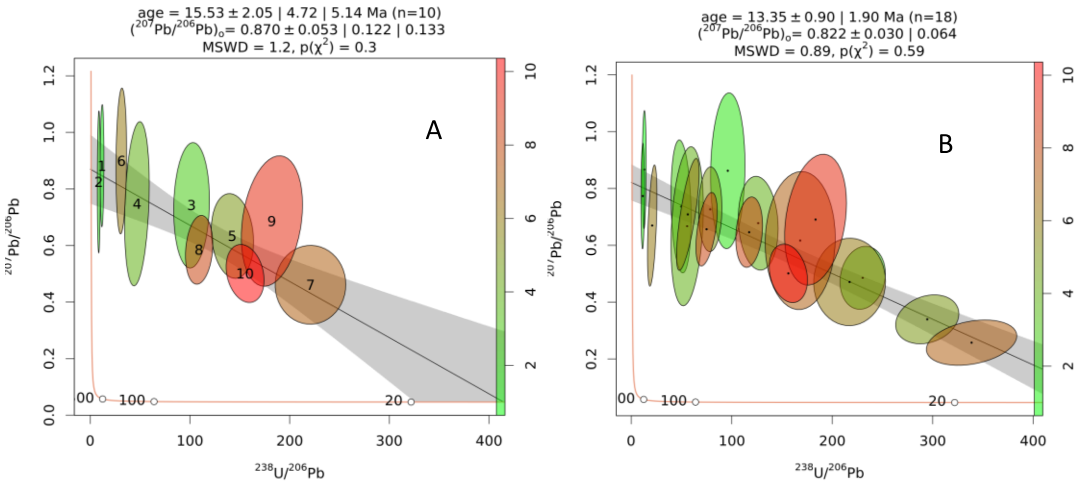

Figure RA2: Comparison of 10 JT ablation once integrated the whole signal (A) and once integrated to get max. spread of Pb/U ratios including multiple integration intervals per signal showing a clear improvement in precision. Colour code is the same ablation have the same colour.

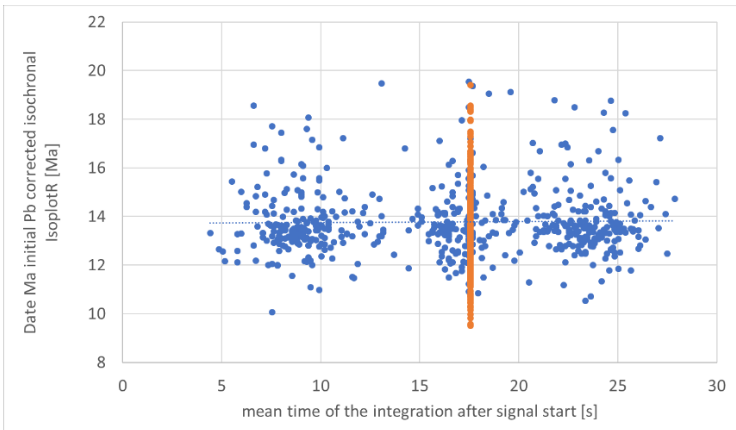

Figure RA3: Individual, isochronal initial Pb corrected JT ages vs. the mean time of integration after laser starts ablating. Blue points indicate selective integrations often the first part (around 8 seconds) and the second half of the signal (around 24 seconds). No significant older ages were found for integrations in the second part. Red points indicate the same for always integrating the whole signal.

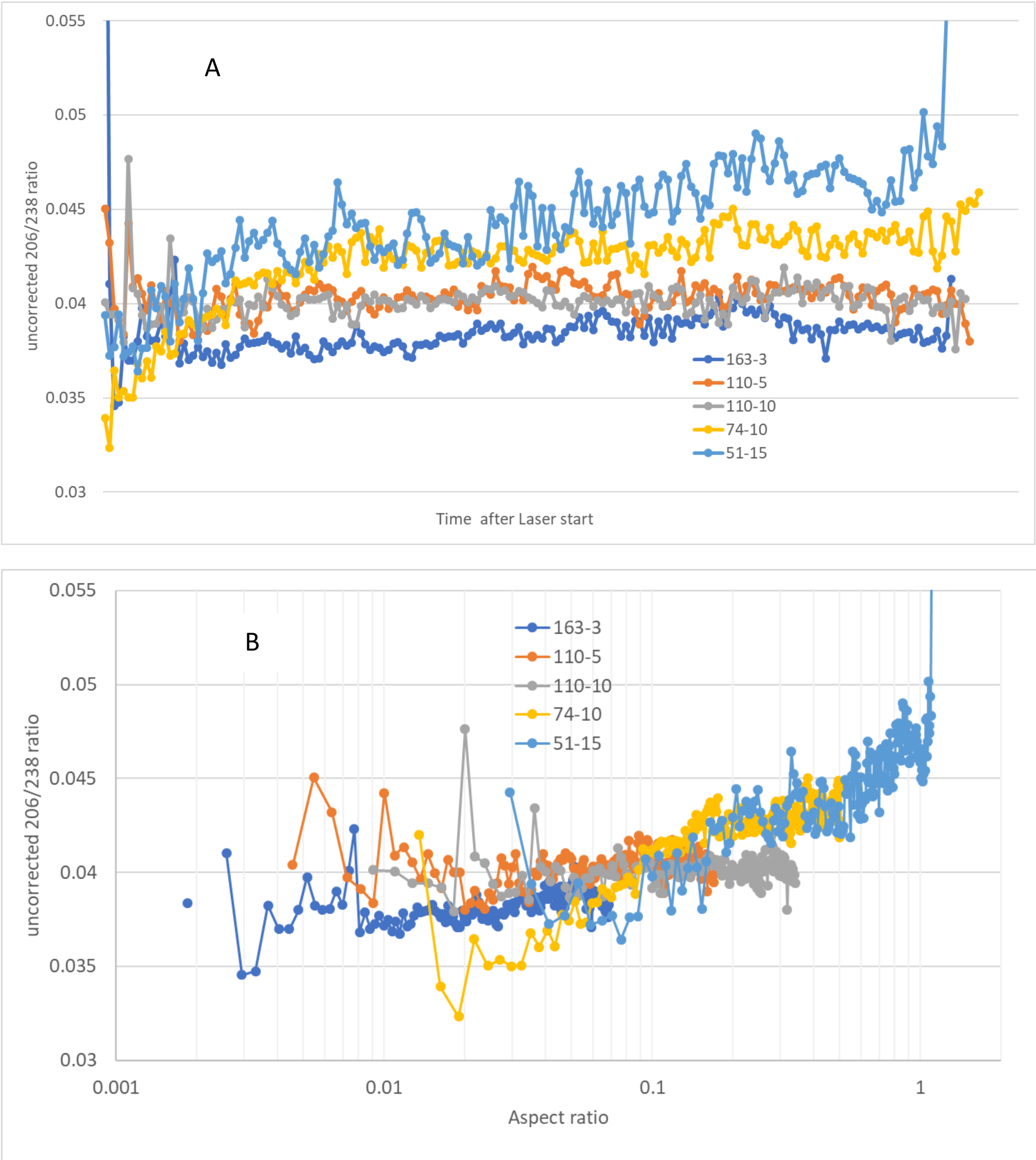

Figure RA4: The downhole fractionation trend of WC-1 for different aspect ratios (mean of n>15 signals, different crater sizes and repetition rate from 163 microns at 3 Hz to 51 microns and 15Hz) both comparing vs time (A) and vs the estimated aspect ratio (B). This figure shows that up to an aspect ratio of 0.2-0.3 no clear downhole trend is observed and only for higher aspect ratios the Pb/U increases significantly.

Answers to RC2:

Please also see also answers to RC 1:

Our method does not use the drift or downhole correction of Iolite. We just use the software to do the baseline subtraction and calculate raw ratios that we subsequently use in an Excel spreadsheet to do the matrix and drift corrections. This is now precised in the text.

“Which bring me on to a question about lines 189-194 – have you tried this?
Yes, this is our standard approach to carbonate dating: We load a bunch of interesting samples, program a short sequence with 2-4 points per interesting phase with shortened gas blank / ablation times and run this “test” sequence (20-30 minutes). We monitor live the U and Pb intensities and decide based on the 2-4 point if the tested sample may work, needs larger crater or is hopeless. Criteria for samples worth a shot are ^{238}U>~10000 cps; or >~1000 cps if there is not much initial Pb, in which case we increase the crater diameter / repetition rate from 110 micron and 5Hz to 163 micron and 7.4 Hz. If ^{238}U < ~1000 cps and/or the initial Pb is higher than ^{238}U, there is little hope, and we skip this sample in the subsequent session.

It is quite common that carbonate samples have a very high MSWD when analysed by ID-TIMS. For instance, the calibration data for WC-1 have a MSWD of 1069 (Roberts et al., 2017). We do explicitly not give the detailed ID-TIMS data for ASH-15 in this manuscript as these data will be published elsewhere (P. Nuriel et al., in preparation)

Thank you for suggesting using the VizualAge_UcomPbine DRS in Iolite, we will give it a try soon.

We adjusted the manuscript with your suggestions. VRM is defined on Line 10 in the abstract.

Answers to SC 1:

Thank you for the comments and suggestions that helped to improve the manuscript. Please find below some answers, comments and findings.

Please also see also answers to RC 1 and 2:

Thank you for pointing out parts of the long and intensively discussed, investigated literature on inter element fractionation. We are well aware of the large amount of interesting and good literature about it, but we would not like to write a review manuscript on this topic so we kept the referencing and discussion to the necessary part, also with respect to the length of the manuscript. Additionally, we investigated downhole fractionation in carbonates, but no clear results emerged (e.g. Figure RA 1) or some investigations using the matrix matched synthetic reference material of pressed powder MACS-3 which is "completely useless for U/Pb dating of carbonates" due to heterogeneity and non-reproducible U-Pb fractionation behaviour.

We added some more detailed investigations on the JT samples and pieces available for distribution, including an image of all available pieces and results of 10 analyses on all possible aliquots that can be distributed. Part of our work over the past couple of years consisted in investigating the suitability of other possible secondary RM, but so far all tested materials but JT were unsuitable.

We added a whole new chapter to the Electronic appendix describing the available pieces of JT in more detail. We agree that JT is both of limited supply and use for low sensitivity ICP-MS, however we think that it is with all the limitations a valuable addition to the collection of possible RM and usable as VRM for some laboratories.

Line 72: We now state:

"The selection of different integration intervals along a single hole ablation can introduce systematic offsets if not randomly distributed due to different amounts of downhole fractionation between RM and sample if there is significant amount of downhole fractionation in either the RM and/or the sample. Best practice is to use as good as possible the same integration intervals with respect to crater shape for both the RM and the sample. As is demonstrated, it is likely that random variability of downhole fractionation, ablation rate, distribution of initial Pb etc. would anyway mitigate the offset potentially introduced. This potentially introduced offset would anyway be diluted in the propagation of the systematic uncertainties, especially since the long term excess variance of secondary RM could precisely result from this."

Line 140-145. See answer to RC1 about this topic.

Section 2.33

1) *U heterogeneity is often at a smaller scale than the pit depth/diameter. Such that a high U zone in the pre-screening may turn out to be low U for the next 30 seconds of ablation (and vice versa).*

We agree, but the opposite is also true: a low U signal during screening may have yielded a higher U signal if ablated longer. This is why we systematically ablate several screening spots per sample, to get an idea of not only the U and initial Pb contents, but also their variability. This fast pre-screening gives some first hints and indications of the possibility if the sample is datable or not, which saves a lot of spot programming and analytical time. From personal communications, we know that other laboratories apply similar strategies.

2) *Chemical and physical heterogeneity can be large, such that the drill rate probably changes a lot during 30 secs of ablation, not just between different materials.*

Yes, we agree, as discussed earlier for replies to other referees. We believe that this is at least partly the reason for the larger excess uncertainties we obtain here for carbonate U-Pb dating, compared to zircon geochronology.

3) *This two-step strategy would add a lot of time to the workflow.*
4) The extra time that this short test sequence costs is much less compared to the time lost if samples with very low U are present and not identified. So we would argue that on the long run, this two-step strategy saves time and resources. *How is the rep rate adjusted to exactly match the aspect ratio? – for a fixed focus point at the surface of the sample, drill rate is non-linear as you drill down such that exact estimation of depth is difficult.*

The rep rate is adjusted assuming equal ablation rate per pulse independent of the crater size and depth. Ablation rate variations due to laser focus are generally small in LA up to an aspect ratio of 1 (i.e. for most aspect ratios considered in this study and carbonate dating in general, for which large spot diameters are used), and likely more influenced by the sample properties (Horn, 2001).

5) *Post ablation measurement of pits will also add significantly more time to the overall data workflow.*

Yes, we do not suggest doing this routinely and this would not be necessary anyway if one follows our suggestion to match aspect ratios between RM and unknowns. The depth measurement was specifically performed here to evaluate the importance of the ablation rate and aspect ratios.

6) *Applying the correction is not actually tested for some unknowns here, so we don't really know if this two-step strategy is going to be an overall improvement for heterogeneous materials.*

We assume that this comment is about the correction for similar aspect ratio, and if so, yes we do not show data for "heterogeneous" unknowns and only for Ash-15D and JT. Of course, we cannot exclude that strong heterogeneity in ablation rates for unknowns would result in age offsets. However, as discussed above (reply to reviewer 1) this effect is likely to be mitigated between different spots and matching as good as possible the aspect ratio of the unknowns to the RM is the best way to minimize it.

7) "A detailed study on how to best apply this correction if necessary is beyond the scope of this work. . ." – it would appear then that this is basically untested. The authors critically undermine *their arguments in their final sentence of this section: "we suggest for a more robust data reduction to always use similar aspect ratios". I would argue that this is exactly what should have been done, and that the community needs to try and find more RMs with a range of U contents.*

What is beyond the scope of this manuscript is to present a method that includes a correction based on crater depth measurements post ablation. What is improving the data quality and versatility is to match the aspect ratio of the pre tested samples to the RM. Using this the amount of ablated material and intensity of U can be adjusted closer to that of the RM. As long as there are not more RM with matching U content and ablation behaviour are available this is in our opinion the best possibility to get best possible precision and accuracy. Additionally, the huge variability of U contents in carbonates, and the fact that we don't really know how high it is a priori is a good argument to for our approach. Matching aspect ratios, rather than spot diameters and U content with a series of RM, is much easier and more cost-/time-efficient at present and is the only option having only one reliable primary RM (WC-1) in the community.

We rephrased to make this point clearer.

Section 2.5:

We rephrased this section to better reflect the information presented in Roberts et al. (2017) concerning the white zone, added the isochrone to the white part and the ages for comparison.

"The plots in Roberts et al. (2017) demonstrate that the white Th-rich region analysed in their data is high in common lead, but that the data-points presented seem to be broadly of the right age. It would appear that the author's data shows that some of these altered regions are not the same age. However, rather than a different age being implied, this is just as likely to be variation due to open-system behaviour; an age might not be definable. So, WC-1 may be heterogeneous in age or homogeneous in age with white zones of alteration causing open system U-Pb behaviour in these zones. When this alteration occurred relative to the dated phase is not resolvable and could have occurred quickly after the formation age or sometime after."

We are convinced that the two phases (the dark and the white zones) show carbonate precipitation at two significantly different points in time / in the burial history of the Capitanian Reef Complex. The black zone (which is the regular part of WC-1 that should be used as RM) is composed of marine botryoidal cements (Roberts et al. 2017) while the white zone is composed of vein-like, more sparry cements (again showing two zonings of different luminosity as can be seen by cathodoluminescence microscopy) with sharp contacts to the surrounding botryoidal cements. If the difference in U-Pb between the botryoidal cements and the more sparry white cements would just be the result of open system behavior, the two cements would not show the difference in texture reflecting different ambient conditions during precipitation that we observe (e.g. crystal size, shape). Also, the U-Pb data of the younger (leached) phase would not define an isochron (which it more or less does, aga 203 +/- 7 Ma) but rather be random distributed towards younger ages. The larger scatter in the data of the white zone can potentially be explained as artefact due to mixture of the dark and white phases in the lower part of the ablation pits (the vein might be tilted) and by the fact that later-diagenetic phases like veins in our experience are commonly more noisy compared to WC-1. We therefore argue that WC-1 is heterogeneous in age. However, the white zones can easily be seen by cathodoluminescence (it has a very bright luminosity compared to the botryoidal cements) so that it is not a problem for the community to avoid the white zones.

Line 241: It is the offset of the average LA-ICP-MS age for ASH15-D compared to the ID-TIMS age. As the ID-TIMS results for ASH15 are not part of this manuscript but in (Nuriel et al., in prep.) and there is an offset observable in both Figure 2b and 4c, we do not ignore this but mention that there is this offset, and a possible explanation is given in the conclusion: "This offset cannot be explained completely by differences in ablation rate and may be an additional matrix effect to be investigated in detail in future work." This offset is already mentioned in the Abstract: "Additionally, a systematic offset to the ID-TIMS age of 2-3% was observed for ASH-15D but not for JT. This offset might be due to different ablation rates of ASH-15D compared to the primary RM or remaining matrix effects, even when chosen aspect ratios are similar."

We consider changing the title as suggested and present the data in a different light.

Attached is a file and table S3 that we would like to add to the electronic supplementary information of the main manuscript containing analyses of all the available parts of JT.

Answer to Editor:

Dear Axel,

Thank you for the additional comments, which I answer below in detail.

"However, two invited experts plus an additional comment have pointed out various points where there is some ambiguity, details need a better explanation or have not been correctly presented. Besides, they have made several suggestions for improvement or clarification. In their reply, the authors have clarified most but not all of these issues mentioning they have fixed this in a revised version. This discussion is in many parts useful for the reader, and I would encourage the authors to implement it as much as possible in the manuscript (or external data depository)."

We implemented as many comments and parts of the answer to the referees as we think is good for the manuscript without going into too much speculation about downhole fractionation. Especially we included the additional measurements on JT on the actual pieces that are available for distribution.

"I agree with most comments/explanations of the authors in discussion with the reviewers and as well with most statements made in the manuscript, but I recommend using careful language and clarifying critical issues, e.g. those express by Nick Roberts."

We think that with respect to the initial characterisation of WC-1 by Nick Roberts we now have clarified a some points so that it is clear that he mentioned "WC-1 is not the perfect material because of its modest heterogeneity". As we see 238/206 ratios ranging from 19 to 34 (almost 50%, compared to the published 25%) and the intercept ages of 203 instead of 254 Ma is about 20% off. We think this is not only a "modest" heterogeneity and likely not only an open system behaviour as the points clearly do not fall on a single mixing line, but rather two. For this we improved Figure 7 now showing two Isochrones.

The text on the WC-1 Heterogeneity is improved:
"We document heterogeneity in the currently most used calcite RM WC-1(Roberts et al., 2017). This information is meant as a caution to analysts to make sure they understand that carbonates are quite heterogeneous materials and are prone to diagenetic alteration including open-system processes at various stages in their geological history resulting in zones with the potential for different ages being

recorded. Care needs to be taken as some WC-1 aliquots may have larger age variation than initially described (Roberts et al., 2017)."

"Similar as the reviewers, I am skeptical about using different ablation conditions between RM and unknowns. The lab in Frankfurt, as an example, has shown that it is possible routinely to use the same spot size, energy density, and frequency between RM and unknowns (perhaps scarifying the spatial resolution in some cases) despite of low or high U concentration. The concept of aspect ratios, as applied in this study, is not common practice at all, and there is no fundamental detail study that has validated it (if I am wrong, please cite it)."

While we understand the skepsis to the presented method and we respect the excellent work done in Frankfurt we would like to point out that it is not only the spatial resolution in some cases that is sacrificed, but the possibility to change the amount of introduced material. There is a well known correlation between sensitivity and precision in ICP-MS. The proposed method, if used carefully and similar aspect ratios are used can be used to date samples that were previously not possible ore only possible with inferior precision using the identical crater size either due to limitations from low count rates or detector cross-calibration issues (practical usable linear dynamic range of the detector). The main focus of the manuscript is not to promote the use of different crater sizes but the observation and warning that in the rare case when using different crater sizes (for whatever reason) without matching the aspect ratio an offset is introduced, so if it is necessary to change the crater size the aspect ratio should be match by changing the repetition rate to improve accuracy.

"The presented data is not entirely convincing; applied to ASH 15D with WC-1 as RM, the ages are 3-5% too young, and in case of JT, ages do not change in the aspect ratio miss-match of 0.8 to 1.7 (due to higher uncertainty, but still its no validation…)."

While we agree that we have an offset for ASH-15D we think that overall, the correlation between aspect ratio offset and age mismatch is quite convincing, even if the uncertainties for ASH and JT are overlapping within the investigated range, the trend is clear for these 2 samples and definitively convincing for WC-1.

Secondly, the authors claim due to different drill rates observed for dolomite (up to 1.6 aspect ratio miss match) and aragonite (1.9, n = 1!) a significant age offsets of 4-8% for dolomite and even higher for aragonite. However, they have not demonstrated it! Besides, the authors not demonstrated that different drill rates using a fixed spot diameter (e.g., for WC-1) yielded the same age offset as by varying both spot diameter and drill rate (see Figure 4). So they link two different observations, which cannot simply be linked without validation. I would be grateful if the authors could add something to the text to make their case stronger. However, I am already happy when they express it in a more careful language and mentioning that further validation is needed.

We agree that we so far did not validate any age offset for dolomite and aragonite due to differences in ablation rate when using identical ablation parameters. Therefore we now use a more careful language only suggesting this effect, pointing out that this needs validation.
"We presume that these differences in ablation rates could result in significant age offsets in a roughly estimated range of 4-8 % for dolomite (160 % ablation rate compared to WC-1)and 6-11 % for aragonite (200 % ablation rate of WC-1) based on the offsets found in figure 4. However, this hypothesis needs validation and any attempt to date dolomite or aragonite needs careful validation."

Please see the revised manuscript with all changes we propose, including a more suitable title:

"Evaluating the reliability of U-Pb LA-ICP-MS carbonate geochronology: matrix issues and a potential calcite validation reference material."