Response to Reviewer 1 (Richard Ketcham) Comment on gchron-2021-22

We thank Richard Ketcham very much for his constructive review. His comments help to improve readability and clarify points that we are trying to make in the text. His questions are insightful and they provide us with the opportunity to elaborate on some important topics that are relevant to the research but hard to deal with in this focused contribution. Although the questions may seem straightforward, the responses are more complicated and must take into account multiple factors. Any additional information that encourages readers to try using a multikinetic approach is most welcome. We reply (blue text) to Richard's comments (black text) below.

Reviewer 1: Richard Ketcham

Multi-kinetic effects in AFT thermochronology have long been neglected by much of the community, I gather in large part because it entails more trouble and expense to acquire sufficient compositional data, and the rewards are unclear, especially since thermal history inversion software will often produce a result without it. Hopefully this paper, and others from this group, will bend the curve.

We understand the reluctance of much of the thermochronology community to acquire elemental data if they are uncertain that the extra cost will yield information that significantly improves data interpretation and modelling. Apatite chemistry, provenance, and thermal history interact in complex ways that result in natural samples that can have a bewildering range of characteristics. We believe that the method we present to define AFT kinetic populations is an improvement over conventional methods but do not claim that it will work for samples that do not exhibit clear multikinetic behaviour. For example, insufficient sample heating may mean that mixed provenance signatures are dominant and therefore incompatible thermal history information is retained within different components of a sample (this was discussed in paper 1, McDannell and Issler 2021, GChron v. 3, 321-335). However, the interpretation of partially annealed AFT samples of variable provenance may still be enhanced with multi-elemental data.

Our observations demonstrate that the rewards of collecting elemental data are worth the extra time and cost because we can better understand the cause of large age dispersion in sedimentary samples that experienced sufficient heating and long residence times in the partial annealing zone and use this information to extract valuable details concerning the thermal history. This improved understanding gives us increased confidence in our model results and new insights into the thermal history of our study areas. Furthermore, one compositionally diverse multikinetic sample has the potential to yield substantially more information than single 'monokinetic' samples, and therefore fewer samples may be needed to address specific scientific questions, which may lower the overall costs of some projects. It should be mentioned that elemental data can be beneficial even in the case of a single AFT age population. We have examples where single populations have high track retentivity based on their elemental composition. Without such data one may assume that they are F-apatite and model them that way. Why does this matter? The absence of elemental data can influence some interpretations. As an example, we have an AFT sample from a thrust sheet in western Canada that has long track lengths and an age somewhat younger than its stratigraphic age. This could be interpreted easily as evidence for thermal resetting and rapid cooling related to thrusting. That would not be correct because the sample is track retentive due to elevated Fe concentration (important for increasing retentivity in the Carlson et al., 1999 dataset) and it is best interpreted as a

volcanically-derived sample that underwent minimal annealing following deposition. Cretaceous volcanism is well documented in the Canadian Cordillera and we have numerous examples of minimally annealed retentive volcanic apatites in Cretaceous rocks of western and northern Canada. One advantage of the LA-ICPMS AFT method is that you can obtain apatite U-Pb ages to check for potential volcanic sources.

At the same time, to be effective in doing so (or at least transparent in trying), it would be good to better document the costs.

We are happy to discuss costs in our reply to the review but we don't think that costs should appear explicitly in the paper for the reasons given below. Certainly costs are one of many factors influencing a research program. However, prices are ephemeral and variable depending on who is doing the work and this not something that is normally included in a scientific paper. The purpose of our paper is to present a different approach to interpreting and modelling AFT data. We believe that the focus should be on the results achieved and not on the added cost of obtaining more data to constrain interpretations. Ultimately it is up to individual scientists to decide how they will allocate their resources. In our case, the benefits of obtaining elemental data are very clear and the adage that you get what you pay for is very applicable. EPMA adds a small premium to the cost but the added value is immeasurable because it enables a reliable differentiation of kinetic populations that may not be possible using conventional methods (Cl, Dpar).

It is not straightforward to predict total lab costs because these will vary among different labs. For example, we are dealing with two laboratories in the United States and costs will be influenced by the Canadian-US exchange rate. There can be separate rates for academic/collaborative research versus commercial work. Some university labs charge significant overhead rates for EPMA which can amplify costs and be a factor in deciding which lab to use. LA-ICPMS AFT analysis can be cheaper and certainly faster than EDM (depending on laboratory) because it avoids sample irradiation, cool down waiting periods, and extra counting of induced tracks. As a result, more grains can be analysed which is necessary for better characterization of different multikinetic populations. In general, AFT costs per sample are fixed, whereas EPMA costs vary with the number of grains and elements being analysed. This will also depend on apatite yield and the number of grains with AFT data that are suitable for probing. Currently we use a standard set of 14 elements that have been observed in variable abundance in Phanerozoic samples from western and northern Canada.

We have a good arrangement between GeoSep Services for LA-ICPMS AFT analysis and the Peter Hooper GeoAnalytical lab at Washington State University for EPMA. Lab costs are reasonable and both labs coordinate their activities. The cost of single-spot elemental analysis per grain is reduced as the number of analyses increases and can vary between US$3.50 (large batch) to US$5.50 (small batch). These costs may be higher if using some other labs that are less specialized for this process than WSU (GeoSep services provides a very efficient framework for selecting points for analysis). For our last contract with WSU, we had 1123 elemental analyses for 18 samples which took 6 days and cost US$3960, coming close to $3.50 per grain with an average cost of $220 per sample. The price of elemental analysis for each sample varies depending apatite recovery. On average, elemental analysis increases costs by approximately 20% relative to just LA-ICPMS AFT, Dpar and apatite U-Pb ages analysis. We believe this cost is easily justified for the complicated, chemically heterogeneous AFT samples we are working with.

For example, how long does the EMPA protocol take per spot?

We consulted with microprobe analyst colleagues, Dr. Owen Neill (formerly of WSU and now Manager, R.B. Mitchell Electron Microbeam Analysis Lab at University of Michigan) and Dr. Scott Boroughs, Peter Hooper Geoanalytical Lab, WSU) and obtained the following information. Probe routines can vary between labs - apatite can be a bit tricky to measure and different labs may optimize for different things, i.e. some labs might optimize for high-accuracy/high-precision F and Cl without bothering with the other trace elements, some labs might optimize for REE's at the cost of some accuracy/precision with the halogens, some might optimize for sulfur, and some labs may have different routines optimized for different things that they alternate between depending on the customer. So going to another lab besides WSU might mean that the probe procedure ends up being entirely different (it might not, but that's something that would be handled on a case-by-case basis). The raw analysis time for our samples at WSU is 3.7 minutes per spot (~16 grains per hour), but that doesn't include setup (1-2 hours), standardization (~8 hours no matter how many grains), and point picking (~150 grains per hour). Another complicating factor is that WSU bills by the hour ($55), but caps at 12 hours a day ($660). The GSC has almost always been able to take advantage of this, with large efficient batches that can be run for 48-72 hours straight, effectively cutting the hourly rate by 30-40%.

What considerations went into decision(s) of whether to do compositional analysis first versus laser ablation? Does doing laser ablation first save time, by figuring out which grains work and providing evidence of whether there is kinetic dispersion, and does this outweigh the disadvantage of not getting the analysis precisely where the tracks were measured?

We thank the reviewer for asking these questions. We present our current approach to doing the analysis but variations on this method can yield similar information and have other advantages. Our process considers and tries to balance: (1) efficiency and speed of analysis, (2) the need to maximize the amount of track length information, (3) minimizing selection bias for age grains, and (4) the desire to obtain replicate elemental data, all as part of routine sample analysis. Samples can be processed faster if all AFT analyses are done first before samples are sent to WSU for elemental analysis. Basically this is an extension of standard methods of AFT acquisition without elemental data. It avoids delays related to transmitting samples back and forth between labs and any delays related to conflicting lab schedules. In principle, doing AFT analyses first could also influence some researchers' decisions on whether to proceed with elemental analysis. For example, if AFT results are consistent with a single age population, some researchers may opt not to do EPMA. In our case, we proceed with elemental analyses for all our samples based on experience. The majority of our samples contain chemically variable detrital apatite with multiple age populations so AFT analysis followed by EPMA is part of our routine work flow. It should be pointed out that elemental data can be useful even for samples with single age populations, especially if track retentivity is much higher than expected for a F-apatite composition (see above discussion on volcanic apatite).

We believe that our procedures work well for many samples. Cf-irradiation is used to increase the number of length measurements in order to reduce the permissible range of thermal solutions that will fit the data during modelling. Length measurements are obtained from apatite grains with and without age information (some grains are suitable for length measurement but not age measurement) to maximize the amount of data. If the number of track lengths is low, then the sample has lower resolution and it becomes difficult to differentiate between simple and more complicated thermal solutions. Abundant length data for our LHA003 sample requires

three heating/cooling events to fit the data (this is observed in other Paleozoic samples in the area) but other samples with far less length data allow for simpler solutions with two cycles. An important issue is bias in age grain selection. LA-ICPMS AFT analysis allows for small grains to be analysed and this is important for proper characterization of single grain age distributions. Generally larger grain sizes may be preferred for EDM work but this type of preference may bias observations, resulting in poor representation of certain apatites or the failure to identify different kinetic populations within a sample. This is especially important when length measurements are not associated with age grains and must be assigned to a population using elemental data and/or other information such as Dpar. Our results from replicate elemental analyses on grains having age and length data indicate that zoning does not appear to be a major problem for many of the Phanerozoic samples we have analysed from northern and western Canada. Results indicate that zoning may contribute to some outlier grains that plot in the "wrong" region of kinetic space based on their age. We do not consider this a serious issue because the amount of kinetic population overlap is substantially reduced when using r_{mr0}/eCl in comparison to Dpar or Cl. We have samples that can show no obvious population overlap or that have some grains that cross kinetic population boundaries within the expected ±0.03 apfu uncertainty range. We have other data from crystalline basement samples with two probe spots per grain, one near the laser ablation pit and the other from a different grain location. These results suggest compositional zoning is minor, but in those few instances, does produce some variation in r_{mr0} values.

Experienced microprobe analysts tell us that acceptable elemental totals are in the range, 97–100 wt %. We are dealing with detrital apatite grains of variable size and, for some samples, we can obtain lower than ideal elemental weight % oxide totals, generally for some of the smaller grains. The reasons for this are variable and sample-dependent (see reply to I. Duddy comment). Elements that were not analysed and the size and physical state of the grain likely play a role. While acquiring data for a variety of Phanerozoic samples from various regions of Canada over a number of years, we continued to refine data acquisition, interpretation methods and modelling procedures based on what we learned. We analysed various suites of elements and some elements were dropped because they occurred in extremely low abundance or were not present. For example, we discovered that elements such as S and Si could occur in high abundance for some samples but these elements were not included in all the analyses because they do not regularly occur in significant abundance nor do they appear as variables in the r_{mr0} model calibration. We do not include these elements in the 'other' category of the r_{mr0} calculation, unless an element was specifically discussed in the original papers. If missing elements contribute to lower totals then the effect should be to slightly increase apfu values for low abundance elements relative to calculations based on ideal elemental totals. The effect would be to slightly increase eCl values relative to those for an ideal wt % total. At some point you need to settle on a suite of the most common elements that have been encountered and this may not include all possible situations.

Another likely cause of lower elemental totals is a reduction in the EPMA measurement area due to imperfections on the polished mineral surface that are related to laser ablation and multiple etching treatments to reveal spontaneous and Cf tracks. Lower elemental totals generally occur with very small grains. Fortunately, a comparison of results for grains with ideal and less than ideal elemental totals shows that similar elemental proportions and r_{mr0} values are obtained for both cases. Therefore, we believe that both sets of data are useful for qualitative binning of apatite grains into different statistical kinetic populations. Due to the statistical nature

of the problem and the uncertainty in absolute kinetics for retentive apatite grains, a pragmatic approach is to use as much of the data as you can to qualitatively assign grains to different kinetic populations and employ the relative annealing approach for modelling that we describe in the paper. We do not believe that avoiding measuring small grains (age bias) or rejecting "less-than-perfect" data in a multi-parameter data set is a better approach. Below we discuss alternative procedures that may improve elemental analyses for small grains. Although less common for our samples, high elemental totals can occur also. This is usually a sign that F/Cl ingrowth or excess oxygen from halogens are not properly dealt with. When analyses show excess halogens outside the limits of apatite stoichiometry, it is not possible to estimate OH content.

Are there cases where changing the order would be a good idea?

Yes. We discuss the procedures we have been using to acquire multikinetic data but the method can be adapted and customized to fit the requirements and preferences of individual labs. There are several factors that can influence whether or not one can obtain a representative elemental analysis. For example, if many of the grains have significant compositional zoning then it would be better to obtain probe data at the point where the age is measured. We have observed cases where replicate analyses with reasonable elemental wt % totals can yield eCl values that vary by > 0.2 apfu. This extreme variation within single grains is uncommon in our samples (usually only in one or two grains in a small number of samples) but has been observed for both EDM and LA-ICP-MS AFT samples that were analysed using different laboratories. There will always be trade-offs because no two multikinetic samples are alike. Generally, you won't know if zoning is a problem in advance of the elemental analysis.

A more common issue than zoning for our samples is related to multiple treatments on the grain mount that can reduce the suitability of smaller grains for probing (see comments above). In some cases, this could be mitigated by using narrower beam widths but it may also mean that an imperfect surface is measured or that no measurement is possible. These conditions may contribute to some reduced elemental totals. In our opinion, the added track length information from Cf-irradiation outweighs the potential lack of probe data for some grains which was the problem for some of the smaller grains in the Permian sample presented in the paper. Choosing to laser the grains as the last step in the process could allow for elemental measurement in the area where spontaneous tracks are counted. Doing elemental analysis prior to Cf-irradiation is not feasible and not cost effective when acquiring lengths off grains that are not used for age analysis. In this case, the results of Cf-irradiation are needed for selecting grains for track length measurement and a second round of elemental analyses would be necessary, adding more steps and reducing the benefits of running samples for EPMA in large batches.

Another beneficial change that could mitigate some of the uncertainty associated with the above issues is to (1) obtain many of the track lengths from age grains, or (2) only use age grains for length measurement. We have done the former for some samples and this can reduce ambiguity in cases where grains with track length measurements lack probe data, have poor probe data, or where zoning is an issue. At least measurements can be linked to age populations rather than solely relying on chemistry. However, this approach does slow down the AFT analysis so it has not been done routinely. We have found that information from replicate grains with both age and length information and two separate elemental analyses have been very helpful with data interpretation when zonation is an issue. In the latter case of only using age grains for length determination, it is possible to do elemental analysis prior to Cf-irradiation

and laser ablation. The success of this approach depends on obtaining enough lengths to provide well constrained thermal histories.

Ultimately there is a balance between efficiency of data acquisition and the amount of effort involved in data reduction and interpretation. The advantage of measuring lengths only on age grains is that probe costs are reduced because you only need EPMA data for up to typically 40 age grains and you can avoid the extra step of identifying grains with replicate analyses. The disadvantage is that you may obtain less track length information because undated grains or those that may be unsuitable for dating can yield abundant length information. Also relocating each age grain for length measurement is slower than just moving across the mount and measuring lengths for selected grains. Whatever the order of operations, we think that it is a good idea to use Cf-irradiation to increase the number of length measurements when track densities and U concentrations are relatively low in order to improve the resolution of model thermal histories.

[line 44] Justifying the 20°C bound seems to require citing Donelick et al (1990), and optionally Tamer and Ketcham (2020).

Thanks for pointing this out. These references have been inserted into the paper.

[line 243] Change to “thermal history modeling” (or “all model calculations”).

The first recommended change has been made to the text.

[line 249] Although replicate values are indeed important for assessing the reproducibility of kinetic parameter values, they may also be taken as an indication of the presence of zoning. The authors do not specify how many spots they took per analysis, but I suspect the answer is one, and that it reflects the usual 2-µm activation zone for EMP; was this driven by the desire for a faster and/or less expensive analysis?

Our investigation of reproducibility of measurements encompasses zoning issues which are discussed in this section and elsewhere in the paper. The cause of divergent measurements may be due to different factors such as poor analyses or zonation. We mention that single spot elemental analyses are used in line 256 in the paragraph immediately below. To further clarify this point, we have modified the sentence on lines 248-250 (new insertion in red font) to be “Replicate elemental (a single EPMA spot per AFT analysis) and Dpar analyses from separate measurements on grains with both age and length data (step 2, Fig. 1) are very important for assessing the reproducibility of kinetic parameter values (Fig. 2).” A single EPMA spot is a very practical choice for minimizing the cost and time for analysis. Our results indicate that it works well for most cases and we feel it is not worth the cost or time to do multiple EPMA measurements to try and eliminate the occasional outlier points that inevitably appear in some data sets. The empirical r_{mr0} model does not incorporate some elements such as S and Si that have been observed in abundance for some samples and thus some scatter can be related to incomplete model calibration. In such cases, additional probing may not help reduce scatter. For the case of difficult samples with lots of zonation, it may be worth redoing the elemental analysis in more detail if it can help improve the interpretation. However, one needs to acquire single spot EPMA data first to determine if more work is required.

Likewise, how many Dpar measurements are averaged for each Dpar determination? The usual procedure is to average four, which ought to make the reproducibility better than observed in Fig. 2c.

The LHA003 sample was analysed at AtoZ Inc. following protocols established by Ray Donelick and these procedures continued to be followed at GeoSep Services. The majority of Dpar measurements were derived by averaging four Dpar measurements per grain analysis. We modified a sentence (lines 152-155, p. 7; new insertion in red font) in section 2.1 to read, "Following standard mineral separation and grain mounting and etching procedures, spontaneous tracks are counted, Dpar is measured for individual apatite age grains (average of four Dpar measurements where possible), and grain x–y coordinates are recorded so that subsequent measurements can be linked to the age grains." Details such as this are also included in the separate GSC open file 8821 that contains the sample AFT and elemental data. Although most Dpar values are based on averaging four measurements it was not always possible to obtain four measurements from every grain. This is particularly true for age measurements where there may not be enough etched fossil tracks to get four values. In contrast, following Cf-irradiation, it is easier to get four values from freshly etched Cf tracks. Therefore, Dpar values associated with age measurements are from spontaneous tracks and Dpar values associated with length measurements are largely from Cf tracks. In principle, this is not supposed to matter because Dpar should depend on the bulk etching properties of the mineral and not on the type of track being used. For the AFT annealing experiments, Dpar measurements were taken from newly formed tracks induced by irradiation.

Also, it's a little unfortunate that the discussion of the downsides of this procedure (lines ~326-340; might not get a compositional analysis near the counting area, or for the grain at all, I gather partly due to the LAICPMS spot) is in the next section; the authors can probably clarify and condense things by briefly mentioning these here, and then referring to them in section 2.3.

On rereading this section of the paper we would prefer to leave it as written unless the Editor prefers that we make this change. It was written this way because we wanted to mention this point in the context of a real example. Single-spot elemental analysis is not a problem for all samples. For example, the LHA003 sample is well behaved in that each grain was probed and it shows minimal population overlap on the age versus eCl plot. Also missing probe data is not an issue for many samples. It happens to be an issue for the P013-12 well sample that has many small grains.

[line 269] Although Dpar imprecision is certainly responsible for a lot of the scatter in Fig. 2e, it's not clear it's the main reason; the authors might try only plotting the points within the 20% bars in Fig. 2c and seeing what the Dpar vs. eDpar scatter looks like. The even scatter might simply be an indication that the things that throw Dpar off are bidirectional; a little OH might increase resistance to annealing compared to no OH (i.e. F-apatite), and a lot of OH might decrease it (e.g., OH-apatite HS from Carlson et al. (1999), but the more OH you have the higher Dpar is.

[Figure 2] Maybe smaller symbols would be better to avoid some of the "solid cloud" effect; some "N =" annotations also would not hurt, and maybe correlation coefficients for d and e.

We have changed the symbols to try and reduce the point saturation. It is tricky to get around this when you are plotting thousands of points at a reduced scale. Given the very large scatter

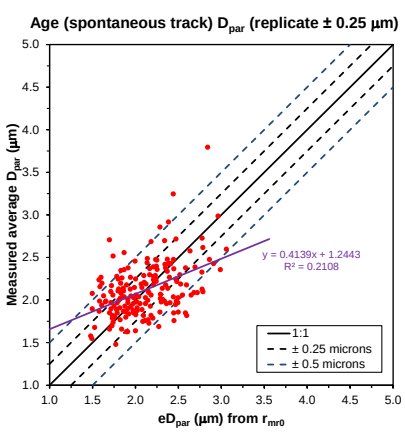
of the data in Figs. 2c, 2d and 2e, we chose not to use linear regression but instead show how the data are distributed around a 1:1 line. We prefer not to include correlation coefficients because we don't think the relationships would be very meaningful, especially for Fig. 2d where there are strong systematic differences between eCl (a function of many elements) and Cl. Coefficients would change with the addition of more data from apatite of different elemental composition because eCl can vary considerably for a given value of Cl. The plot data are included in the supplement to the paper should someone wish to analyse it in a different way. The number of data points is shown in the title for each plot so adding "N=" would duplicate information that is already there. We have not made this change but if the Editor prefers this alternative we can remove the information from the title and add the notation.

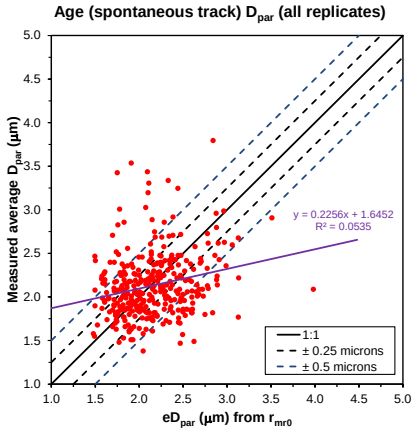
We think that figure 2c and 2e are a fair representation of the uncertainty involving Dpar measurements when applied to natural samples and, from an applications viewpoint, it is unclear to us what can be gained by focusing on a subset of the data with less scatter. When dealing with natural samples, one does not have the luxury of only choosing data that appears to be "well-behaved." Nevertheless, we made some plots below of measured Dpar versus eDpar for various combinations of the grains with replicate analyses to illustrate our point. For some samples, Dpar may show low variability on a plot like Figure 2c, but there are two separate issues here. One is the ability to accurately repeat measurements on the same grain (Fig. 2c) and the other is to be able to resolve kinetic populations with these measurements (Figs. 4 and 5). One might expect reasonable reproducibility of Dpar measurements for the single crystals with induced tracks that have been used in annealing experiments. However, it would appear that Dpar measurements for detrital apatite grains show considerably larger uncertainty.

Figure 2c shows replicate Dpar analyses on the same apatite grain by the same analyst. It is difficult to separate out all the factors that may influence Dpar measurements such as grain chemistry, grain/track orientation, use of variably annealed spontaneous tracks versus fresh tracks formed by irradiation, the number of measured tracks for obtaining an average Dpar value, and analytical issues (etching, measurement, etc.). In principle, one might expect to see much better reproducibility in the measurements if the analyst revisited each mount and measured the same etch pits. However, in our case, the etch pits associated with the age measurements are from spontaneous tracks and those associated with length measurements are from Cf tracks. If much of the scatter is related to differences in etching between fossil and freshly formed tracks then this should have implications for using Dpar as a kinetic parameter because it is supposed to depend on mineral solubility and not the origin of the track being measured. We have noticed that plots of Dpar versus eCl show higher correlation coefficients for Cf tracks than for spontaneous tracks for sedimentary samples from the Mackenzie Delta region of northern Canada. If etching behaviour is the same for both types of tracks then other factors may be contributing to variability in Dpar measurements between age grain and length grain measurements. Under these conditions, four measurements may be insufficient to produce a representative average value. Although it is reasonable that compositional variation between grains within a sample can cause variation in Dpar values, it is less clear how variation in composition within individual grains with replicate analyses can lead to the large data scatter in Fig. 2c.

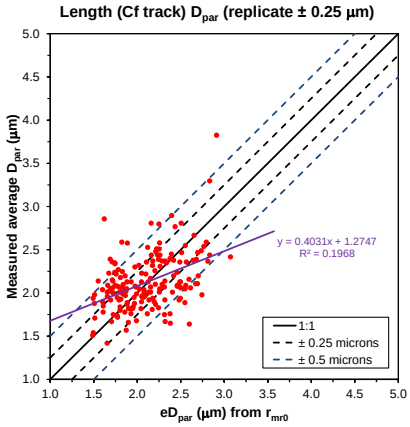
The two plots below are for Dpar measurements taken from age grains with replicate data that were used in Fig. 2c. The eDpar values were converted from elemental data gathered on the

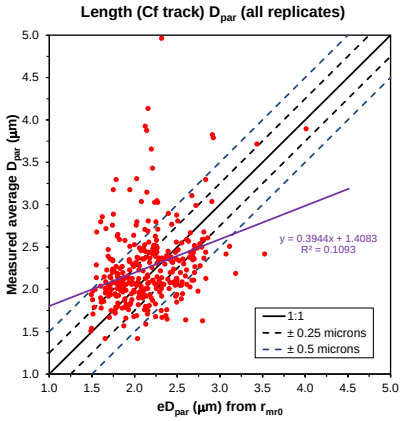
age grains. The left panel is for grains plotting within the 0.25 μm contours in Fig. 2c and the right panel is for all the age grains in Fig. 2c that have elemental data.





Next we show the same type of plots but for Dpar values associated with length measurements taken after Cf-irradiation. The eDpar values are from elemental data gathered during probing of grains with length measurements.





We can see that both the age Dpar values (spontaneous tracks) and length Dpar values (Cf tracks) show very similar results (left panels above) for the grains with replicate Dpar analyses that are within ±0.25 μm in Fig. 2c. However, both the age and length data show significant scatter with respect to eDpar. This scatter is consistent with our observation that r_{mr0}/eCl/eDpar are much better at resolving kinetic populations than Dpar. The main problem is that many samples have a rather narrow range of Dpar values (~1.0 to 1.5 μm spread). Therefore population overlap is inevitable if uncertainties in measured Dpar are on the order of 0.25 to 0.5 μm. Measurement precision is simply too low to clearly resolve kinetic populations. The left panels are the optimistic case for the better fitting subset of data. Expanding the data to all the replicate samples (right panels) or to the thousands of analyses in Fig. 2e greatly increases the scatter and uncertainty in measured Dpar relative to eDpar.

Regardless of the cause of the scatter, we maintain that Dpar is a low resolution parameter in general. We have a small number samples where Dpar resolves different age populations but this is not the norm. Similarly, we have samples where measured Cl can resolve kinetic populations. For these simple cases, either Cl is the only significant element or Cl covaries

systematically with the other elements which makes it appear as if Cl is the only element controlling annealing. If these conditions are not present, then populations overlap with respect to measured Cl. Obviously the mentioned Cl cases will go unnoticed or there will be a level of 'positive reinforcement' regarding the importance of Cl, if only Cl is measured or considered (see Green and Duddy CC1 and CC3 comments). The whole reason for exploring r_{mr0} in the first place was because a significant amount of multikinetic data from northern Canada could not be interpreted using Dpar or measured Cl.

[line 291] "colour-coded"

Quotations have been added.

[line 295] It may be worth noting that compositional populations may also be good candidates for shared inheritance. Although eCl is one such possibility, insofar as it combines a number of compositional variables into one number, apatites with similar eCl may get there via different compositional components, and thus not constitute a good candidate for shared inheritance. This is discussed further below.

This is an important point but a somewhat complicated and sample-dependent issue that we don't think should be discussed at this point in the paper because it doesn't apply to these samples which behave as kinetic populations without variable provenance signatures. We do touch on this aspect in the discussion. You can't properly assess the role of provenance until after you undertake thermal history modelling. If you can model all kinetic populations together using the same history then variable provenance is not an issue that will significantly affect model results. Either the sample thermal history has erased much of the provenance record or differences in provenance are too small to worry about. We are trying to keep the discussion focused and relevant to the data at hand. We have seen many different multikinetic samples and there are lots of different situations that could be discussed but we want to avoid going off on tangents. Our view is that AFT kinetic populations are dynamic and dependent on mineral chemistry and thermal history. Differential annealing and thermal resetting could homogenize formerly different provenance groups that share similar kinetics. Independent thermal maturation data show that this has happened for a large number of our samples. The elemental data we use to define the kinetic behaviour may not correlate with groups based on provenance information provided by other parameters such as U-Pb ages, REEs or other unique signatures. The relation between kinetic populations and populations based on provenance signatures is an interesting topic for future research as more detailed multi-elemental and age data are acquired. Some provenance information may be preserved throughout the thermal history but kinetic populations may change.

[line 333-337] Maybe here or elsewhere, discuss the choice between switching which bin a grain is in, versus leaving the grain out altogether.

Usually only a small percentage of AFT ages appear as outliers in kinetic space unless a sample shows significant elemental zonation. As we mentioned in the paper, our view is that age data are preferred over eCl data if there is a conflict—because laser spots coincide with track count areas, and EPMA spots depend on finding a 'clean' spot surface to get a good measurement. If ages have relatively high precision and plot well within one of the kinetic populations defined on the radial plot then we rely on the age information to determine which population it belongs to. Under these circumstances, age dispersion can be substantially reduced in one population by moving the grain to the other population while not significantly

affecting the age of the population to which the grain was reassigned. We believe this inference is easily justified for grains with replicate analyses where elemental analyses associated with the age cause the grain to be an outlier in kinetic space whereas the elemental analysis data associated with the length measurement for the same grain moves it to a population with similar ages. Other data can be used to help with interpretations. For example, we have some samples where U-Pb ages are distinct enough to identify a volcanic component as being separate from a detrital component and this can help if the volcanic component constitutes a separate kinetic population. Volcanic apatites can have a distinct and sometimes unusual chemistry that makes them stand out. It seems reasonable to use as much of the data as possible because age grains can have associated length information.

[line 438] The claim that population 3 has retained tracks from 540 Ma, or from about 245°C (Figure 6) is eye-catching, and probably overly optimistic about the ability of AFT to retain information about such high temperatures. It appears to stem from a difference in how AFTINV evaluates total annealing versus HeFTy's "oldest track". HeFTy assumes total annealing after reduced mean length falls below 0.4095 for non-projected lengths, corresponding to a mean length of just under 7 µm, whereas AFTINV appears to have total annealing correspond to a mean track length of 2 µm (line 419). This may be based on a slight misinterpretation of what's written in Ketcham et al. (2000); the 2 µm limit mentioned there corresponds to the smallest track that can appear in a track length distribution. However, such occurrences are due to including a population of tracks with a higher mean and large standard deviation. The 0.4095 value arises in part from the observation that no annealing experiments reported by Green (1988) or Carlson et al. (1999) had a mean length below 7 µm (although there are some 6's and 5's reported by Barbarand et al. (2003), and even an occasional 4 or 3 by their Analyst 3). Willett (1997) uses a similar value of 0.428 as the zero-density intercept of reduced length versus density reported by Green (1988). In other words, by the time a mean length falls below some limit, the track population becomes undetectable. I believe this provides a more realistic basis for evaluating total annealing and the oldest retained track. Using the revised criterion, the TA for the oldest track for an rmr0=0.491 apatite is closer to 200°C, which seems a lot more reasonable considering the closure temperature is 161°C. This is not the most crucial of issues, but it's prudent to avoid distracting claims.

This is a fair point to make. We realize that our choice of shorter track length for modelling retention ages gives an uppermost maximum limit on temperatures for track retention and we agree with the reviewer that we should not put too much credence in such theoretically-derived temperatures. This value was chosen because it represents the shortest track length ever measured by Ray Donelick out of many tens of thousands of analyses. We have some samples with measured track lengths between 2 and 4 µm but these are rare. The reduced mean length of totally annealed tracks is easy to change for the AFTINV model but doing so won't change the concepts illustrated by the theoretical retention ages described in the paper. We will modify the abstract to remove references to specific model-dependent annealing temperatures and update the text to better explain this point. From the reviewer's comments we can see that this value is not a precise parameter and it depends on our ability to observe and measure short tracks. As discussed by Ketcham et al. (1999), the concept of annealing temperature depends on how it is defined and it is influenced by factors such as annealing kinetics and the heating and cooling rates. However, the concept is still useful for estimating when track information may be retained in a sample. The retention age calculations still show how various populations have been annealed and they give an approximate oldest possible time limit from which AFT

populations can start to retain thermal history information. This is important for our samples because it shows that provenance information has been erased for some of the kinetic populations and therefore it is possible to model all the data using the same thermal history.

[Figure 6, 7] I appreciate the authors' efforts to incorporate the CRS method into AFTINV, and intrigued by the result – it looks to be a powerful addition. I have long been considering doing something similar myself, having dropped the CRS method when I converted my earlier program AFTSolve to HeFTy. However, one of the reasons I did so may still be evident in the model results here. The CRS method has a tendency to quickly converge to a relatively smooth solution that does not explore the solution space as well as the Monte Carlo method, and thus map out the range of solutions that fit well. In HeFTy results, this allows the resolving power of the data to be evaluated by looking at the width of the solution envelopes.

We thank the reviewer for drawing attention to this important point on clarifying the meaning of the thermal history results and the use of the CRS algorithm. The algorithm uses an expansion factor, alpha, that controls how aggressively the algorithm searches for new solutions by randomly recombining members of a fixed set of model histories to generate a new trial history. Earlier versions of AFTINV included a different implementation of the CRS algorithm but we stopped using it for the same reasons given by the reviewer. A critical consideration is that, previously, a constant alpha was used and the initial starting pool of random solutions were not acceptable solutions. Therefore, the starting point was far from acceptable solution space, it could take a long time to find solutions, and the algorithm could get stuck in a minimum and generate too narrow a region of acceptable solution space.

Two important changes were made. First, the initial pool of solutions is generated using the random Monte Carlo method and it converges when all solutions exceed the 0.05 significance level. The CRS algorithm is now able to draw from a set of acceptable 0.05 solutions and tries to improve the entire set to a higher significance level. The second change was to cycle through different values of alpha in order to help prevent the algorithm from getting stuck. If no solutions are forthcoming at a selected value of alpha, then alpha ramps up. When a solution is found, alpha drops and the cycle may be repeated. A more aggressive search is beneficial earlier in the calculations but it is not beneficial to stay at a high value of alpha. As the model evolves and generates many more solutions, then the range of alpha values decreases. There is a balance in selecting alpha values and iteration times. The algorithm we settled on is based on testing different schemes using different multikinetic samples to determine which gave the overall best performance in terms of maximizing the number of 0.5 solutions within reasonable model run times.

We believe that this new version of the CRS algorithm is a powerful addition to AFTINV because it enables us to discover the conditions required to obtain close fitting solutions. If variable provenance is not a factor, then multikinetic populations add very powerful constraints on thermal histories due to the requirement to fit the AFT age and length data of all populations with the same history. There are multiple reasons why it can be difficult to obtain answers. It is easy to determine if variable provenance is a factor by modelling individual populations and comparing the thermal histories. If pre-depositional thermal histories do not overlap, then this suggests different provenance information is preserved. For our samples, the main difficulty in finding solutions was to ensure we had the appropriate style of thermal history and compatible kinetic parameters. In the case of our LHA003 sample, three cycles of heating/cooling were required to fit the data. Early models had only two cycles and they were unsuccessful. Also, the

ability to closely fit age and length data depends on having the right separation in kinetic parameter values and this can be estimated using the CRS calculations. The number of 0.5 level CRS solutions is related to the relative differences in eCl values between different kinetic populations so we can determine the range in offset that produces the most close-fitting solutions.

We view the CRS smoothing of thermal solutions as a desirable effect for the temperature ranges that the AFT data are sensitive to, especially since the resulting CRS history pool is derived from the starting pool of histories that already fit at the 0.05 level. Unlike HeFTy and QTQt, AFTINV constructs thermal histories differently and uses many more time points on a fixed, user-specified grid. The Monte Carlo calculations use uniform random deviates to generate new temperature points. This creates more possibilities for small fluctuations and stair-step patterns in the thermal solutions. There is a trade-off between time and temperature so it is easy to overshoot or undershoot temperature due to this compensation effect. These fluctuations are not resolvable by the data and ragged looking thermal histories can yield just as good fits as smooth histories. We prefer the “averaging effects” of combining solutions to yield less extreme heating/cooling rates at the 0.5 level due to data resolution issues. We try to have the best of both worlds here. The 0.05 level Monte Carlo solutions map out a broad range of acceptable solutions and this envelope can be interpreted the same way as it is for the HeFTy model. The CRS solutions are then embedded in this envelope to show close-fitting, smooth solutions at the 0.5 level. The Monte Carlo solutions allow higher rates and therefore upper temperature limits and average temperatures are somewhat higher than the CRS values. For LHA003, thirteen 0.5 solutions generated by the Monte Carlo algorithm are more ragged looking and extend beyond the range of the CRS solutions because of that.

Although the CRS algorithm tends to yield smoother solutions, it can converge on multimodal solutions if they are permitted by the data. For P013-12, there is a minor solution mode with older peaks. The low number of track lengths in population one allows for a broader range of thermal peaks for the second heating/cooling event. The solutions with older Jurassic thermal peaks are associated with Cretaceous thermal peaks for the second heating event. Sample LHA003 also shows more than one solution mode. The number of permitted modes depends on the amount of length data available to constrain solutions and the complexity of the thermal history. We have samples with fewer tracks for which the CRS algorithm yields multimodal solutions. This indicates that the new version of the algorithm is less likely to be trapped in a thermal minimum, since multiple modes at the 0.5 significance level indicate no clear ‘preference,’ or alternatively a general inability to update the solution set and therefore preferentially select one of the modes as a result. The exponential mean solution generally provides an excellent fit to the data if there is a dominant solution mode. However it may fail if there are different modal peaks.

In the results here, what puzzles me for P013-12 is the relatively tight band of good solutions above 175°C from 600-450 Ma, and probably a fair bit younger/cooler than that. Given the 161°C closure temperature of the most resistant population, the idea that it would exert much constraint in the 175-250°C temperature range seems improbable, and is not reflected in the QTQt results either. This all is not necessarily a problem, but I think it should be discussed so people interpreting these results have a more complete knowledge of what they are looking at.

This is a good point to make. The model temperature at which tracks are retained will be a function of the annealing kinetics defined by eCl and the rate of temperature change and so

there will be a range of model temperatures corresponding to the estimated times of total track annealing. For example, very rapid heating (on the order of 10°C/My) following deposition for the LHA003 samples means that tracks will survive to higher temperatures than for more modest heating rates. This variability in rates and temperatures is reflected in the distribution of model retention ages so it is difficult to assign a single temperature for assumed total annealing. Using a higher value of reduced track length (~ 0.4) to represent total annealing will result in younger retention ages but the same concept applies. The ability of AFT data to constrain thermal histories diminishes as temperatures approach total annealing temperatures and the only way to estimate temperature ranges where AFT data can constrain the thermal history is to examine the model behaviour.

For the Devonian sample, LHA003, the most track retentive population has an eCl value of 0.5 apfu. The exponential mean solution shows a shift to more steady cooling rates at ~480 Ma and ~175°C. The average cooling rate for the exponential mean solution below 175°C is ~1.4°C/My between ~480-380 Ma. Above ~175°C, the cooling rate drops off. For the Permian sample, P013-12, the most retentive population has an eCl value of 0.55 apfu and steady cooling rates are achieved below a temperature of ~185°C at ~440 Ma. The average cooling rate below ~185°C is ~1.2°C/My from 440-295 Ma. Above ~185°C, the cooling rates decrease significantly. We concur with the reviewer that the AFT data do not constrain the thermal history at temperatures > ~175-185°C. Our interpretation is that the change in cooling rate marks the approximate point at which AFT data are able to constrain the temperature history. So why are the CRS temperature envelopes so narrow between 175-185°C and 250°C? This is related to the nature of the calculations. CRS solutions are required converge on a narrow temperature interval of 245-250°C, the same starting condition for the Monte Carlo calculations. These high temperatures were chosen to ensure that all populations were totally annealed at the beginning of the thermal history. Rate and temperature boundary conditions are enforced during CRS calculations to maintain consistency with the Monte Carlo calculations and solutions that fall outside of these ranges are not accepted. This tends to limit the spread of solutions in the region where AFT data provide no constraint. Probably the best way to deal with this is to annotate the figure to indicate where the thermal history is unconstrained by AFT data.

We believe the AFT data constrain temperatures at less than ~175-185°C for the samples used in the study. At these lower temperatures, each population retains some record of the thermal history. The lowest retentivity population is most sensitive to the later, lower temperature portion of the history whereas higher retentivity populations provide more constraints for the earlier higher temperature part. Thermal history resolution is enhanced because preserved tracks in each population contribute information on overlapping portions of the thermal history.

Along similar lines, did both the AFTINV and QTQt models assume that all apatites in each sample had the same inherited, pre-depositional history? If so, was the fact that they did so, and their success in fitting their models, and indication that there was shared provenance, or an indication that, for these samples, results are not terribly sensitive to the pre-depositional history?

We cannot determine from the available data whether each kinetic population has the same shared inheritance because this information has been degraded by thermal annealing. AFTINV modelling was done before QTQt modelling. Individual populations were run first to get an initial impression of the thermal history. For LHA003, the pre-depositional thermal record for the lower retentivity population one was erased completely by thermal annealing. Therefore, the pre-

depositional history reflects information retained in the high retentivity population two and only this population is sensitive to this inherited history. This enabled us to run both populations together in order to enhance the resolution of the post-depositional thermal history which is the main focus of our northern Yukon study. It is notable that LHA003 shows more variability in the pre-depositional cooling history than P013-12 and this may be related to a higher degree of annealing associated with the higher temperatures (>30°C) encountered during the first thermal peak.

For P013-12, the pre-depositional thermal history has been erased for population 1 and model pre-depositional thermal histories overlap for population 2 and 3. Population 2 has also experienced significant annealing so the pre-depositional history is dominated by population 3. Again it is unclear whether there was shared provenance because thermal annealing has obscured the information. It should be noted that our multikinetic scheme can still be useful for interpreting and modelling samples where some populations preserve different pre-depositional histories. For example, we have a sample with four different populations, three of which can be modelled together. The fourth population was much older with a different pre-depositional history and was modelled on its own. The r_{mr0} parameter still allowed us to define separate kinetic populations. Within QTQt, we did allow a variable pre-depositional history during modelling for both samples. This is apparent in the broad t-T envelopes for the Expected model. The results indicate that due to the strong annealing, the pre-depositional history does not greatly influence results, especially for sample LHA003.

Or, are the results sensitive – do the few earlier-cooling 0.5 paths for P013-12 corresponds to the earlier peaks T's at ~195 Ma and/or ~70 Ma?

The four earlier 0.5 level cooling paths correspond to a cluster of solutions with temperature peaks in the range, 141-144°C at a model time of 172.5 Ma. This is a solution mode determined by the CRS algorithm but it is doesn't seem that there is much sensitivity to the pre-depositional cooling in terms of affecting later thermal events. The two outlier paths with temperatures in the range, 145-148°C at 192.5 Ma have associated thermal peaks at ~70 Ma but their pre-depositional cooling paths are within the region defined by the majority of CRS cooling paths. Therefore, for this sample, a pre-depositional history can be resolved but it has little impact subsequent thermal events.

The manual (AFTINV) and automatic (QTQt) raising of the rmr0 values for the most resistant populations in each sample is interesting. What seems to be going on is that the different populations need greater separation in their partial annealing zones to produce their respective divergent age and length distributions. It's further interesting that the higher resistance is corroborated by the vitrinite data for sample LHA003, though less so for P013-12. The authors recommended approach of "anchoring" on low-resistance kinetic seems like a good one. Another possible "advantage" of the Ketcham et al. (1999) model over the (2007) one beyond the different rmr0 equation is that it has a much higher temperature range, which these results may imply is necessary to create these divergent populations.

Yes we believe that relative annealing is a powerful approach to modelling multikinetic AFT samples when we lack the data to accurately predict kinetic parameters for some of the apatite populations that are likely to be encountered in nature. It takes advantage of the following: (1) the r_{mr0} parameter is calibrated based on the relative annealing behaviour among apatites of different composition in the annealing experiments; (2) observed AFT data for different kinetic

populations within a sample must be accounted for within a shared post-depositional thermal history framework; (3) lower retentivity apatites within a specific compositional range are abundant in nature and are best represented in the annealing experiments and therefore provide a reasonable reference point for estimating kinetic parameters for less well understood compositions; and (4) thermal maturity data provide independent paleotemperature information for assessing kinetic parameter assignments and model thermal history predictions. Organic maturity is consistently high within Paleozoic strata across the northern Yukon study area and compatible with the eCl values required for the higher retentivity AFT kinetic populations. For the P013-12 sample, a single measured maturity value from a Permian cuttings sample is uncertain. It is similar to the overlying Cretaceous section and may be biased to a lower value. The Permian is only ~90 m thick in this well and the underlying Carboniferous units with more measurements have maturity values that are higher by ~0.1 %Ro. The model still fits the Permian measurements within two standard deviations.

Lastly, the comparison between AFTINV and QTQt results appears to gloss over their differences a bit. For P013-12, the first reheating peaks at ~168 Ma in AFTINV and could go as far back as 195 Ma, whereas QTQt appears to strongly say that it was at about 140 Ma. Similarly, AFTINV implies that the first peak reheating for LHA003 was at 345 Ma, compared to 300 Ma for QTQt. If you lay the models pairs on top of each other, they appear to exclude each other at these times. Is this because QTQt calculated different kinetics than the manually-shifted ones in AFTINV, or because of QTQt favoring simpler histories, or some combination of these and possibly other factors?

One of the main reasons that we ran the QTQt model was to show that the general style of thermal history determined for our samples using AFTINV could be reproduced using a different modelling approach. People who are not used to working with multikinetic data may be skeptical of the detailed thermal histories that we obtained for our samples. Prior to modelling the samples, we had a poor understanding of the regional thermal history. Much of the stratigraphic record has been eroded from the study region and simple model boundary conditions based on two major thermal events failed to yield satisfactory solutions for the LHA003 sample. Therefore, modelling investigations were undertaken to discover the conditions required for obtaining successful solutions. The approach was not to force any preconceived notion of the thermal history but to learn from the data. The QTQt results confirm that the multikinetic data retain a record of multiple heating events. AFTINV requires that the number of thermal events be specified by allowing for model inflection points to occur over specific time-temperature ranges. If the events aren't required then inflections will be minor and contribute nothing significant to the thermal history. QTQt was run with wide open boundary conditions and minimal constraints. It converged on a solution with three heating-cooling cycles for LHA003 without that condition being imposed.

AFTINV and QTQt models are constructed differently and calculations are undertaken in a significantly different way. Some differences are:

(1) The statistical methods are different. AFTINV uses a "frequentist" approach where the objective is to find a set of thermal solutions that exceed an absolute goodness-of-fit probability threshold. QTQt use a "Bayesian" approach that yields single model solutions (e.g. maximum likelihood or maximum posterior) from a large model ensemble that is a result of maximizing the probability (i.e., likelihood ratio) as a measure of fit to the data, or the 'acceptance' criterion. Therefore AFTINV (like HeFTy) relies on absolute values of

probability, whereas QTQt relies on the (likelihood) probability ratio. QTQt has a different acceptance criterion and lacks a specific 'threshold' for model path acceptance.

(2) AFTINV uses a grid with many time-temperature points whereas QTQt tends to use only the minimum number of points needed to fit the data at an adequate level (as modelled in our examples). The model adds and subtracts points as needed and tries to avoid too much complexity. Although AFTINV uses fixed time points, it imposes heating and cooling constraints consistent with specified thermal history styles to avoid saw tooth temperature fluctuations in thermal histories. QTQt is more tolerant of wide open boundary conditions, in part because it uses a minimal number of time-temperature points. For AFTINV, boundary conditions need to be more carefully specified. If they are too wide open, the much larger number of model points creates many more possibilities which could cause the model to spend too much time interrogating unpromising regions of solution space.

(3) Initial model conditions can vary. In typical applications of AFTINV for sedimentary samples, model thermal histories are started at high temperature at times much earlier than the depositional age. The goal is to try and model any inherited history as a pre-depositional cooling event. This allows the model to cool below the total annealing temperature when required by the data. If there is no inherited history, then a broad range of cooling curves will be generated. For QTQt applications, boundary conditions can be wide-open and it can generate a simple pre-depositional cooling history in an appropriate region of time-temperature space. Therefore, we specified a 'common' pre-depositional history in AFTINV (after trials suggested this was suitable), whereas we allowed different, albeit grossly simplified, pre-depositional histories for each kinetic population in QTQt (i.e., a single t-T point added prior to the depositional age).

(4) For AFTINV, kinetic parameters can be adjusted manually relative to a population with fixed kinetics to maximize the number of close-fitting solutions. In QTQt, kinetic parameters are adjusted automatically within ranges to improve model fits. The program may stop adjusting the parameters if it decides incremental changes in fitting the data are not significant enough. Thus, with many parameters allowed to vary (higher degrees of freedom), QTQt may provide marginal fits to ages or lengths, but those results are conditional on the input data and model prior assumptions—where in this case we rejected 'more complex' models for equivalent likelihood—which essentially provides a lower limit on the t-T complexity required to fit the data (possibly at the expense of 'perfect' fits to the data).

(5) AFTINV generates a set of acceptable (0.05 level) and good (0.5 level) solutions (typically 300 each). The AFTINV solution set maps out a broad range of possible times and temperatures for thermal peaks. The exponential mean of the CRS solutions generally provides a smooth, good-fitting solution. QTQt generates several types of single t-T solutions based on likelihood maximization and the posterior probability. The 95% credible interval around the expected model for example, summarizes (weighted mean or mode of the posterior peaks) all of the accepted post-burn solutions and often denotes regions of better or worse t-T resolution.

(6) Both models incorporate a relation between initial track length and kinetic parameters but they are different. QTQt uses relations published in Carlson et al. (1999) and AFTINV uses a newer unpublished relation provided by Ray Donelick that contains more data.

(7) QTQt specifies AFT central ages as input but converts LA-ICPMS data to EDM equivalent data. Although central ages are used for plotting model results, Ns and Ni count values are used in the QTQt calculations rather than ages. AFTINV uses central ages when there is significant age dispersion but otherwise uses pooled ages. It can model either EDM or LA-ICPMS AFT data and does not do conversion between these two different types of data.

The above factors can account for why there are differences in detail between the two models. For P013-12, AFTINV provides a broader range of good-fitting (0.5) CRS solutions with peak times extending from 150 Ma to ~195 Ma. The QTQt expected model is closer to this lower age limit but the 95% confidence region overlaps with AFTINV results. It is clear that QTQt does not provide as close a fit as AFTINV to the length data for population three. QTQt converged on eCl = 0.47 apfu for population three whereas AFTINV used 0.55 apfu. The most likely reason for the difference in model results are that AFTINV generates more solutions, uses more model points to construct thermal histories, and that observed ages are different for each model due to the conversion of LA-ICPMS FT data into EDM FT data by QTQt. This results in younger ages for the QTQt model because EDM and LA-ICPMS central ages are not equivalent. The difference between QTQt and AFTINV results for LHA003 is due to a number factors as well. These include that manual fine tuning of kinetic parameters allows AFTINV to more closely fit population two ages and lengths. QTQt converged on an eCl value of 0.44 whereas AFTINV used 0.5 which generated 300 0.5 level solutions. The difference in the age of the first peak is likely related to the difference in kinetics and differences in the pre-depositional cooling history. AFTINV shows cooling from higher temperatures later in the pre-depositional history. QTQt does prefer simpler histories and uses fewer points so that is definitely a factor as well.