



# Reproducibility of Thermal History Reconstruction From Apatite Fission-Track and (U-Th)/He Data

Richard A. Ketcham, Peter van Der Beek, Jocelyn Barbarand, Matthias Bernet, Cécile Gautheron

## ► To cite this version:

Richard A. Ketcham, Peter van Der Beek, Jocelyn Barbarand, Matthias Bernet, Cécile Gautheron. Reproducibility of Thermal History Reconstruction From Apatite Fission-Track and (U-Th)/He Data. *Geochemistry, Geophysics, Geosystems*, 2018, 19, pp.2411-2436. 10.1029/2018GC007555 . insu-03595904

HAL Id: insu-03595904

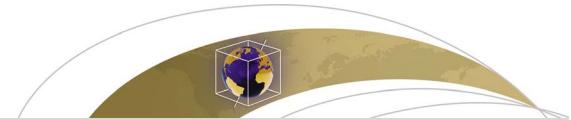
<https://insu.hal.science/insu-03595904>

Submitted on 3 Mar 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

Copyright

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



# Geochemistry, Geophysics, Geosystems

## RESEARCH ARTICLE

10.1029/2018GC007555

### Key Points:

- An interlaboratory exercise examined the reproducibility of apatite fission-track and (U-Th)/He data and thermal history analysis
- Thermal history models agreed on major features but varied on detail, due to differences in data, data entry, model setup, and software
- Although cautionary in many respects, study results point to promising avenues to improve the reproducibility of thermal history modeling

### Supporting Information:

- Supporting Information S1

### Correspondence to:

R. A. Ketcham,  
ketcham@jsg.utexas.edu

### Citation:

Ketcham, R. A., van der Beek, P., Barbarand, J., Bernet, M., & Gautheron, C. (2018). Reproducibility of thermal history reconstruction from apatite fission-track and (U-Th)/He data. *Geochemistry, Geophysics, Geosystems*, 19, 2411–2436. <https://doi.org/10.1029/2018GC007555>

Received 13 MAR 2018

Accepted 3 JUL 2018

Accepted article online 28 JUL 2018

Published online 9 AUG 2018

## Reproducibility of Thermal History Reconstruction From Apatite Fission-Track and (U-Th)/He Data

Richard A. Ketcham<sup>1</sup> , Peter van der Beek<sup>2</sup> , Jocelyn Barbarand<sup>3</sup> , Matthias Bernet<sup>2</sup> , and Cécile Gautheron<sup>3</sup> 

<sup>1</sup>Department of Geological Sciences, Jackson School of Geosciences, University of Texas at Austin, Austin, TX, USA, <sup>2</sup>Institut des Sciences de la Terre, Université Grenoble Alpes, Grenoble, France, <sup>3</sup>GEOPS, Université Paris-Sud, CNRS, Université Paris-Saclay, Orsay, France

**Abstract** We report a new interlaboratory exercise to evaluate the reproducibility of apatite fission-track (AFT) and (U-Th)/He (AHe) data and thermal history analysis. Twelve laboratory groups participated, analyzing apatite separates from two previously studied localities. Ten groups returned AFT data from 13 analysts, five groups returned AHe data, one contributed apatite U/Pb data, and nine contributed thermal history models. Submitted AFT age data were generally consistent with the original studies and each other to within uncertainties, although there were departures, particularly among results obtained using laser ablation inductively coupled plasma mass spectrometry for uranium determination. AFT-confined track-length data showed more variation, which correlated between samples, suggesting that they may reflect analyst-specific factors. Accounting for anisotropy using *c* axis projection reduced the variation and correlation. AHe ages showed more dispersion than observed in the previous studies, with one sample showing many ages consistent with prior work but also many significantly older ages, and the other roughly matching prior work but showing symmetrical scatter suggesting a 1SE uncertainty of 17–21%. Thermal history models generally agreed in their major features but showed considerable variation in detail, due to differences in data and data entry, model setup, and modeling software approach. Addressing pitfalls in data entry and model setup improved congruence, as would greater emphasis on AFT length and etch figure calibration. Because of scatter, AHe data had to be entered selectively into the models to achieve reasonable results. Although cautionary in several respects, study results point to promising avenues for improving the reproducibility of thermal history modeling.

## 1. Introduction

Thermal history inversion using low-temperature thermochronological data is a widely used technique with a range of geologic applications, from rifting and basin evolution to mountain building and landscape development (e.g., Gallagher et al., 1998; Lisker et al., 2009; Malusa & Fitzgerald, 2018; Reiners & Brandon, 2006). The reproducibility of these methods is critically important to their soundness and the degree of confidence that can be placed in their results. While individuals and research groups establish best practices internally, reproducibility is ultimately necessary at the community level for a technique to flourish.

The principal studies for evaluating community-level reproducibility in low-temperature thermochronology have concerned the apatite fission-track (AFT) system. Early interlaboratory experiments compared fission-track age determinations (Miller et al., 1985, 1990; Naeser et al., 1981). For fission-track lengths, Miller et al. (1993) and Ketcham et al. (2015) reported results from exercises in which apatite aliquots were sent out to multiple research groups for length measurements using their respective in-house procedures. A smaller-scale experiment (Sobel & Seward, 2010) looked at the influence of etching conditions on AFT etch-pit diameter ( $D_{\text{par}}$ ). An additional study (Ketcham et al., 2009) examined the reproducibility of length measurements of multiple analysts inspecting the same grain mounts with the same microscope, but also went on to examine how the differences observed affected thermal history inversion. In all of these studies, differences among analysts and research groups were significantly in excess of statistical expectation, and point to the importance of calibration and cross-calibration with age standards and the measurements that underlie the annealing models used to extract thermal histories (Barbarand, Carter, et al., 2003; Barbarand, Hurford, et al., 2003; Carlson et al., 1999; Ketcham et al., 2007b).

To build upon these results, a new interlaboratory study was designed to provide a more comprehensive test of the reproducibility of thermal history analysis, encompassing both AFT and apatite (U-Th)/He (AHe) data, together with the inverse modeling process. In preparation for the 14th International Conference on Thermochronology, held in 2014 in Chamonix, France, two previously studied localities were resampled to provide material, aliquots of which were handed out at the conference. Participants were asked to provide both analytical results and thermal history inversions, using the procedures and facilities at their respective institutions and the modeling software of their choice.

## 2. Experiment Design

Two localities (Bergell massif, Swiss Alps and Mont Lozère in the southeastern Massif Central, France) were chosen based on their characteristic time-temperature histories and expected apatite yield and quality. Both had provided relatively straightforward data in previous studies, and we sought to avoid complications such as variability in provenance or in probable annealing or diffusion kinetics beyond factors that are part of currently routine or widespread analysis (i.e., AHe grain size and effective uranium content,  $eU$ ; AFT etch-figure size,  $D_{par}$ ). Whereas the Bergell massif shows a relatively simple Neogene exhumation history, the southeastern Massif Central is characterized by a more complex post-Paleozoic thermal history including a reburial episode (see below).

Sample S1 is from the Variscan Mont Lozère granite ( $^{40}\text{Ar}/^{39}\text{Ar}$  and U/Pb age  $\sim 306$  Ma, Brichau et al., 2008) located at the southeastern border of the Massif Central near Génolhac, France; it corresponds to sample 19 from Barbarand et al. (2001) and MC19 from Gautheron et al. (2009). The locality features a nonconformity indicating that the sample was near the surface in the Hettangian (earliest Jurassic; 199–201 Ma) prior to reburial. Peak burial was reached in the Early Cretaceous and is interpreted to correspond to the deposition of a kilometer-scale Jurassic and Lower Cretaceous sedimentary section. There is also evidence for widespread late Jurassic (146–156 Ma) fluid activity in the region (Cathelineau et al., 2012). Subsequent basin inversion led to erosion of the section and exhumation of the granite. AFT single-grain ages reported from S1 passed the  $\chi^2$  test for age homogeneity (Barbarand et al., 2001) and AHe analyses using multigrain aliquots reproduced to within analytical uncertainties (Gautheron et al., 2009). No detailed thermal-history modeling has been done at the site of this sample, although modeling results for nearby samples using Monte Trax (Gallagher, 1995) were reported by Barbarand et al. (2001). The AFT and AHe ages from the area lie on fairly linear and parallel, though noisy, age-elevation trends (Barbarand et al., 2001; Gautheron et al., 2009), implying fairly steady exhumation until 40 Ma but not ruling out rate changes, and with little constraint on the final unroofing history. Following sampling, the new material was reanalyzed for AFT to compare to the previous results and provide information that was not measured in the original study (track angles to  $c$  axis,  $D_{par}$ ).

Sample S2 is a granodiorite from the Bergell pluton in the southeast central Alps (Switzerland) and from close to the locality of sample BG27 of Mahéo et al. (2013). Following intrusion at 31–28 Ma (Oberli et al., 2004), the Bergell massif was exhumed during northward indentation of the Adriatic plate into the Austro-alpine Penninic wedge. Previous work utilizing AFT and AHe age-elevation profiles encompassing multiple samples and 1-D thermal modeling suggests that the region exhumed rapidly from 20–25 to  $\sim 17$  Ma, after which exhumation slowed until  $\sim 10$  Ma and was quiescent until 5–6 Ma, followed by another pulse of more rapid exhumation (Mahéo et al., 2013). The inferred exhumation history is principally based on a break-in-slope in the AHe age-elevation profile, which indicates a slowing of cooling between 15 and 17 Ma, with a minimum AHe age of  $5.4 \pm 0.3$  Ma at the valley bottom requiring subsequent more rapid cooling so as to be exposed at present. The modeled profile ends about 3 km from the site of BG27. As with S1, the published AHe ages used multigrain aliquots. There are no AFT data associated with the BG27 locality, although Wagner et al. (1979) reported AFT ages for the nearby age-elevation profile discussed above which also indicate deceleration of cooling in the mid-Miocene.

The original localities were resampled, and standard rock-crushing and mineral-separation procedures were used to obtain apatite separates. These were divided into 30 roughly equal aliquots for distribution to study participants. The only information provided to participants concerning the samples is that both were from granites; that sample S1 was covered by Hettangian (201–199 Ma) strata in depositional, not intrusive, contact; and that the present-day surface temperature is 0–10 °C at both localities. Participants were also given a questionnaire to fill out concerning their analytical and calibration procedures and inverse

**Table 1**  
Summary of Data and Inverse Models Returned for Interlaboratory Study

Lab #	AFT	Data submitted		Modeling software used	
		AHe	Other	HeFTy	QTQt
1		Y			
2	Y			Y	
3	Y				
4	Y		U/Pb, chemistry		Y
5	Y			Y	
6	Y	Y		Y	
7	Y			Y	
8	Y	Y		Y	
9	Y	Y		Y	Y
10	Y			Y	
11		Y			
12	Y			Y	

Note. AFT = apatite fission-track; AHe = apatite (U-Th)/He.

modeling setup. To avoid discouraging participation in this and future exercises, the identity of all participants has been kept confidential. Abbreviations used in this paper are based on sample number (S1 or S2), lab group number (L1, L2, etc.), and analyst within each lab group (A1, A2, etc.). Lab groups were requested to report data in time to be analyzed and presented for the 2016 international thermochronology meeting (Ketcham et al., 2016); to obtain more results, two lab groups that did not attend the 2016 meeting were allowed to submit data in the following months.

### 3. Results

#### 3.1. Participation

Of the 30 lab groups given aliquots for this study, 12 returned data (Table 1). Ten lab groups provided AFT data from 13 analysts. Five lab groups provided AHe data, three of which provided both AFT and AHe data. One lab additionally provided apatite U/Pb data. Nine labs returned inverse modeling results, eight of which used HeFTy (Ketcham, 2005), three used QTQt (Gallagher, 2012), and two used both. Three lab groups provided only data, without modeling.

#### 3.2. AFT Data

AFT data are provided in Tables 2–4 and shown in Figures 1 and 2. Table 2 lists questionnaire results pertaining to AFT analyses for those labs that did them, and Tables 3 and 4 provide the data for samples S1 and S2, respectively.

##### 3.2.1. Ages

All ages are presented as pooled ages and discussed with their 95% confidence intervals. In the case of individual laboratory results, the intervals are as reported by the respective participants, who varied in whether they reported symmetric or asymmetric errors (Galbraith & Laslett, 1985). Dispersion was calculated according to Vermeesch (2017).

AFT ages for sample S1 (Table 3; Figure 1a) range from  $81.9_{-9.7}^{+11.0}$  to  $114.3 \pm 8.0$  Ma ( $N = 13$ ), with a mean of  $97.0 \pm 5.5$  Ma and median of  $94.1_{-8.2}^{+9.0}$  Ma (L4 A1). The mean and median are more consistent with the reanalysis ( $90.2 \pm 9.2$  Ma) than with the originally reported age ( $109.5 \pm 10$  Ma; Barbarand et al., 2001). Overall agreement among analyses was good: 9 of 13 analyses included the median in their 95% confidence intervals, and three of the four others (as well as three of the nine) overlap with the published age. Ages for sample S2 (Table 4; Figure 1b) range from  $13.4 \pm 1.2$  to  $22.4 \pm 2.6$  Ma ( $N = 12$ ), with a mean of  $17.9 \pm 2.1$  Ma and median of  $17.9 \pm 1.7$  Ma. Agreement was again good, with 9 of 12 analyses having confidence intervals overlapping the median value.

Overall, there is no pattern of one lab producing systematically older or younger ages compared to its peers. However, there is an interesting correlation of  $\chi^2$  probabilities between samples (Figure 1c). In particular, analyses using the external detector method (EDM) tended to pass the  $\chi^2$  test, while those using laser ablation inductively coupled plasma mass spectrometry (LA-ICP-MS) were more likely to fail or have low values. Failure to pass  $\chi^2$  was reasonably correlated with divergence; of five analyses across both samples with  $p(\chi^2)$  values below 0.05 (20% of analyses), three (60%) did not have the median age in their confidence interval (42% of divergent ages). Taken as a set, the AFT ages pass the  $\chi^2$  test, with probabilities of 0.37 and 0.05 for S1 and S2, respectively; if the analyses utilizing LA-ICP-MS are removed, the  $p(\chi^2)$  values rise to 0.65 and 0.51.

##### 3.2.2. Lengths and Etch Figures

For sample S1, the median mean track length (L2 A1) is in near-exact agreement with the published value of  $12.70 \pm 0.16$   $\mu\text{m}$  (Figure 2a).

**Table 2**  
Apatite Fission-Track Analysis Parameters

Lab #	Etch <sup>a</sup>	TINTs only? <sup>b</sup>	System <sup>c</sup>	Mag <sup>d</sup>	Length calibration <sup>e</sup>	Age type <sup>f</sup>
2	5	Y	FTStage	1,250	KCH2	EDM
3	5	n.r.	n.r.	n.r.	n.r.	LAICPMS
4	5.5	N	Autoscan	1,000	None	LAICPMS
5	5.5	Y	Custom	1,000	None	EDM
6	5	Y	FTStage	1,000	DUR	EDM
7	5	Y	FTStage	1,250	None	EDM
8	5.5	Y	Autoscan	1,250	KCH2	LAICPMS
9	5	Y	Autoscan	8,000	KCH2	LAICPMS
10	5	Y	Custom	1,250	None	EDM
12	5.5	Y	FTStage	1,250	KCH2	EDM

Note. EDM = external detector method; n.r. = information not reported.

<sup>a</sup>Molar strength of  $\text{HNO}_3$  used for etching. <sup>b</sup>Whether only track-in-track confined tracks were measured. <sup>c</sup>Fission-track measurement system used. <sup>d</sup>Typical microscopic magnification used. <sup>e</sup>Confined fission-track length calibration: KCH2 = participated in Ketcham et al. (2015) exercise, in which aliquot 2 consisted of unannealed Durango apatite induced tracks; DUR = other Durango apatite material. <sup>f</sup>Whether EDM or LAICPMS was used for U determination.

**Table 3**  
*Apatite Fission-Track Data for Sample S1*

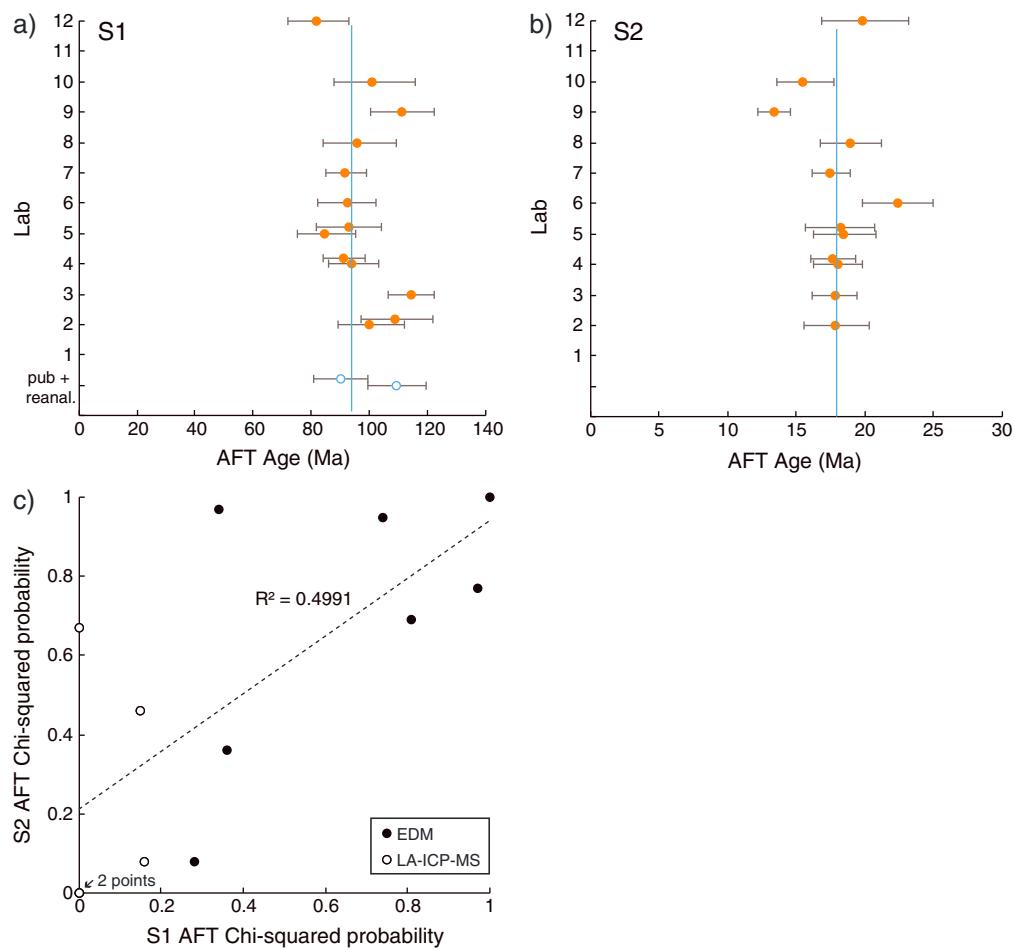
Lab #	Analyst	Pooled age (Ma) <sup>a</sup>	95% conf. int. (Ma) <sup>b</sup>	N <sub>age</sub> <sup>c</sup>	p( $\chi^2$ ) <sup>d</sup>	Disp. (%) <sup>e</sup>	$l_m$ <sup>f</sup> ( $\mu\text{m}$ )	$\sigma_l$ <sup>f</sup> ( $\mu\text{m}$ )	$l_{c,\text{mod}}$ <sup>g</sup> ( $\mu\text{m}$ )	$\sigma_{c,\text{mod}}$ <sup>g</sup> ( $\mu\text{m}$ )	N <sub>l</sub> <sup>h</sup>	D <sub>par</sub> <sup>i</sup> ( $\mu\text{m}$ )	$\sigma_{D\text{par}}$ <sup>i</sup> ( $\mu\text{m}$ )
2	1	100	+12.1/-10.8	20	0.81	2.1	12.7	1.88	13.81	1.24	107	1.74	0.25
	2	109	+13/-12	20	0.99	0	12.73	1.3	13.72	1.1	100	1.45	0.17
3	1	114.3	+8/-8	45	0	17	-	-	-	-	-	-	-
4	1	94.1	+9/-8.2	27	0.16	10	12.01	2.08	13.65	1.27	101	-	-
	2	91.2	+7.4/-6.9	27	0.15	10	-	-	-	-	-	-	-
5	1	84.6	+10.5/-9.3	20	0.36	4.9	12.41	1.49	13.99	0.97	100	2.31	0.38
	2	93	+11.2/-11.2	20	0.74	0.5	12.23	1.26	13.7	1.13	102	2.79	0.25
6	1	92.4	+10/-10	22	0.97	0	12.58	1.65	13.86	1.01	100	1.49	0.15
7	1	91.6	+7.3/-6.7	20	0.34	3.5	14.01	1.32	14.86	0.95	100	2.03	-
8	1	95.9	+13.3/-11.7	21	0	27	12.61	1.11	13.91	0.73	77	1.34	0.05
9	2	111.4	+10.8/-10.8	40	0	22.7	12.75	1.53	13.87	1.11	58	1.31	0.08
10	1	101	+15/-13	20	0.28	8.9	13.54	1.41	14.45	1.01	78	1.69	0.12
12	1	81.9	+11/-9.7	20	1	0	13.56	1.24	14.31	0.97	53	1.78	0.1
Barbarand et al. (2001)		109.5	+10/-10	25	0.30	5	12.7	1.55	-	-	101	-	-
Reanalysis		90.2	+9.2/-9.2	24	0.43	6.3	12.89	1.21	13.96	0.89	87	1.59	0.05
#		13					11		11			10	
Mean		97					12.83		14.01			1.79	
Median		94.1					12.70		13.87			1.72	
Minimum		81.9					12.01		13.65			1.31	
Maximum		114.3					14.01		14.86			2.79	
St. Dev.		9.9					0.61		0.38			0.47	

<sup>a</sup>Pooled fission-track age. <sup>b</sup>95% confidence interval for pooled age, as reported by each participant. <sup>c</sup>Number of grains measured for age. <sup>d</sup>Probability of obtaining higher  $\chi^2$  value from a single population. <sup>e</sup>Dispersion. <sup>f</sup>Mean and standard deviation of nonprojected confined track lengths. <sup>g</sup>Mean and standard deviation of c axis-projected confined track lengths. <sup>h</sup>Number of confined track lengths. <sup>i</sup>Mean and standard deviation of diameter of etch figures parallel to c axis.

However, variation is higher than for the ages. Three of 11 analyses had confidence intervals for mean track length that did not overlap with the 95% confidence interval for the median analysis, and 5 of 11 did not include the median value in their confidence intervals. The result is similar for sample S2 (Figure 2b); although 10 of 11 analyses had confidence intervals that overlap the median confidence interval (L5 A1;  $13.09 \pm 0.18 \mu\text{m}$ ), only 4 of 11 had the median value in their confidence intervals.

**Table 4**  
*Apatite Fission-Track Data for Sample S2*

Lab #	Analyst	Pooled age (Ma)	95% conf. int. (Ma)	N <sub>age</sub>	p( $\chi^2$ )	Disp. (%)	$l_m$ ( $\mu\text{m}$ )	$\sigma_l$ ( $\mu\text{m}$ )	$l_{c,\text{mod}}$ ( $\mu\text{m}$ )	$\sigma_{c,\text{mod}}$ ( $\mu\text{m}$ )	N <sub>l</sub>	D <sub>par</sub> ( $\mu\text{m}$ )	$\sigma_{D\text{par}}$ ( $\mu\text{m}$ )
2	1	17.8	+2.5/-2.2	20	0.69	2.3	13.39	1.59	14.31	1.11	80	1.91	0.22
	1	17.8	+1.6/-1.6	28	0.67	0	-	-	-	-	-	-	-
4	1	18.0	+1.8/-1.7	30	0.08	13	12.63	1.54	13.99	1.04	101	-	-
	2	17.6	+1.7/-1.5	28	0.46	0	12.45	1.76	13.94	1.1	101	-	-
5	1	18.4	+2.4/-2.1	20	0.36	1.0	13.09	1.8	14.28	1.25	101	3.07	0.44
	2	18.2	+2.5/-2.5	20	0.95	0	12.3	1.58	13.70	1.13	102	2.97	0.63
6	1	22.4	+2.6/-2.6	22	0.77	0.2	12.9	1.53	14.02	1.2	87	1.69	0.17
7	1	17.5	+1.4/-1.3	20	0.97	0	13.61	1.52	14.46	1.12	100	2.17	-
8	1	18.9	+2.3/-2.1	29	0	26	13.13	0.99	14.36	0.84	60	1.75	0.03
9	2	13.4	+1.2/-1.2	34	0	13	12.52	2.05	13.83	1.22	88	1.55	0.18
10	1	15.5	+2.2/-1.9	20	0.08	11	13.63	1.46	14.52	1.06	100	1.82	0.12
12	1	19.8	+3.4/-2.9	20	1	0	13.71	1.37	14.51	1	100	2.03	0.17
#		12					11		11			9	
Mean		17.9					13.03		14.17			2.11	
Median		17.9					13.09		14.28			1.91	
Minimum		13.4					12.30		13.70			1.55	
Maximum		22.4					13.71		14.52			3.07	
St. Dev.		2.2					0.51		0.29			0.55	



**Figure 1.** Apatite fission-track ages from S1 (a) and S2 (b). Error bars show 95% confidence intervals. Vertical line is median age value and blue points are from Barbarand et al. (2001) and a reanalysis by JB. (c) Correlation of  $p(\chi^2)$  values for the two samples; two points overlap at (0,0). AFT = apatite fission-track; EDM = external detector method.

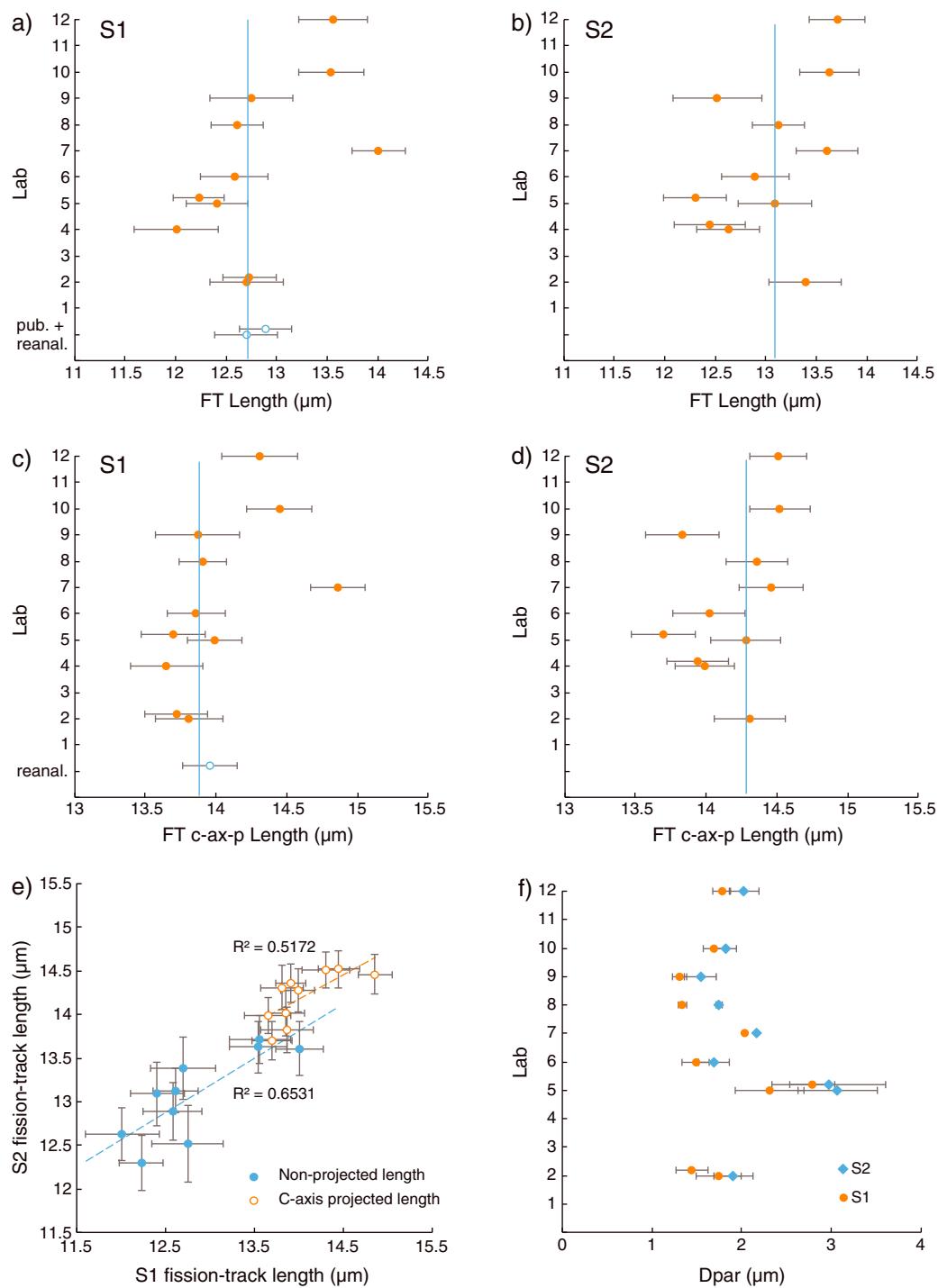
Accounting for track angle using  $c$  axis projection (Donelick et al., 1999; Ketcham et al., 2007a) improved overall congruence. For S1, 8 of 11 analyses include the median of the mean projected lengths (L9 A1;  $13.84 \pm 0.15 \mu\text{m}$ ) in their confidence intervals (Figure 2c). For S2, again only 4 of 11 analyses have the median value in their 95% confidence intervals, but four additional samples had confidence intervals brought to within  $0.1 \mu\text{m}$  of the median (L5 A1;  $14.28 \pm 0.13 \mu\text{m}$ ), despite  $c$  axis projection shrinking the confidence intervals (Figure 2d). The range of maximum–minimum mean value falls from 2.0 to  $1.2 \mu\text{m}$  for S1 with  $c$  axis projection and from 1.4 to  $0.8 \mu\text{m}$  for S2; both drops in range exceed the reduction of errors.

Track-length measurements between samples are correlated (Figure 2e); analysts measuring longer lengths for S1 tended to also report longer lengths for S2, with a correlation coefficient  $r$  of 0.81. The degree of correlation falls somewhat when lengths are  $c$  axis-projected ( $r = 0.72$ ), but is still significant.

$D_{\text{par}}$  values varied widely among nine analysts (Figure 2f). However, ordering is consistent: in all cases, the mean  $D_{\text{par}}$  value for S1 was less than that for S2. This relationship is consistent with the Cl content data obtained by L4 (A1 averaged 0.10 wt% for S1 versus 0.36 wt% for S2; A2 averaged 0.09 and 0.42 wt%, respectively).

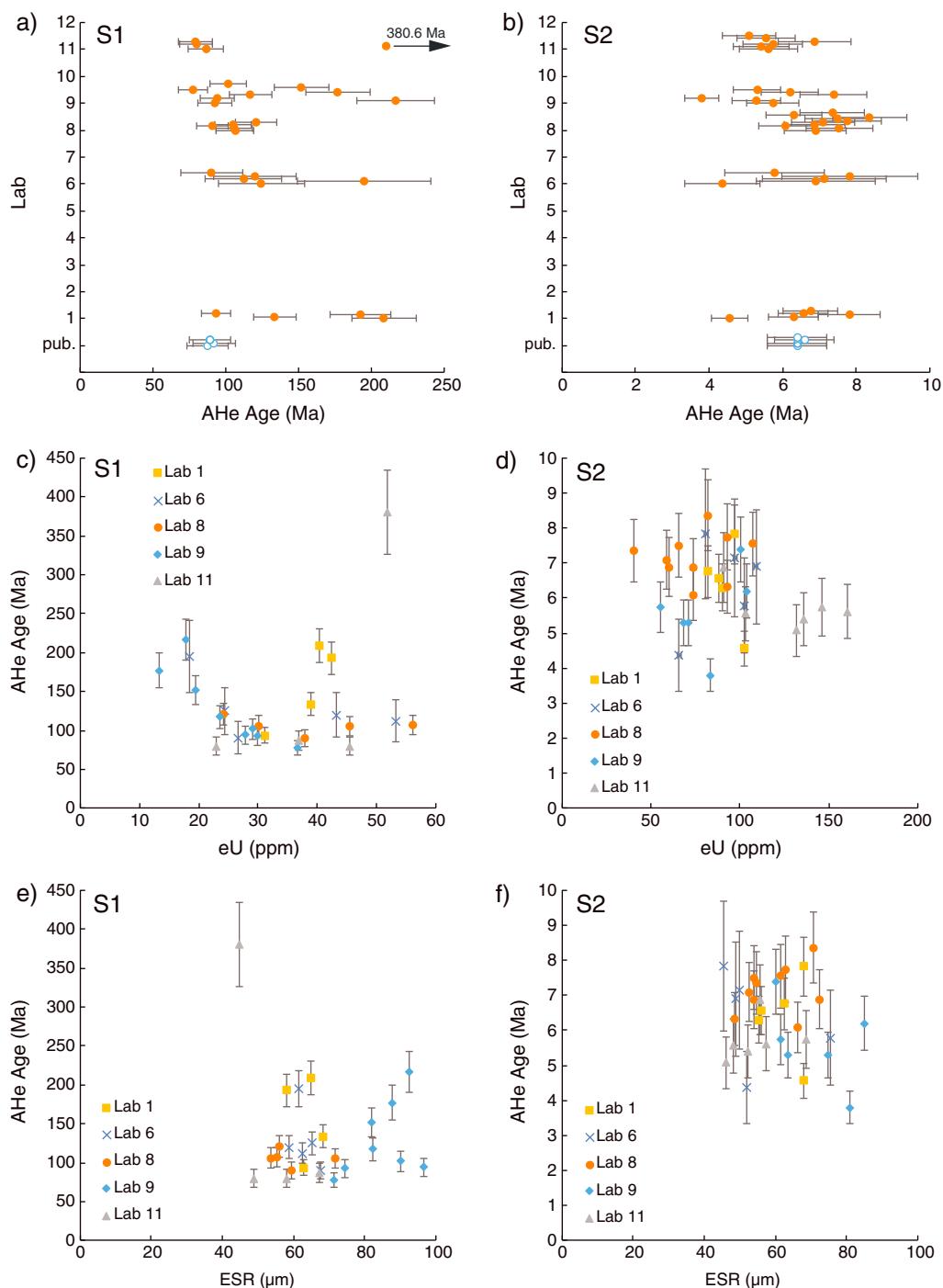
### 3.3. AHe Data

The contributed AHe data are provided in Table 3. All data are as reported by study participants, except where conversions were necessary to put everything into the same table (e.g., ncc to pmol/g  ${}^4\text{He}$ ; ppm  ${}^{147}\text{Sm}$  to ppm Sm).



**Figure 2.** (a–e) Apatite fission-track mean length data; error bars are 95% confidence intervals of the mean. Vertical lines in (a–d) denote group medians. (a) S1, nonprojected lengths; (b) S2, nonprojected lengths; (c) S1, *c* axis projected lengths; (d) S2, *c* axis projected lengths; (e) correlation between length measurements in S1 and S2. (f)  $D_{\text{par}}$  (diameter of etch figures parallel to *c* axis) for the two samples, with error bars showing two standard deviations. FT = fission-track.

Single-grain AHe ages for S1 (5 labs, 26 grains in total) range from 77.7 to 380.6 Ma (Figure 3a, Table 5). All labs had some single-grain ages that were in agreement with the original study ( $90 \pm 4$  Ma), but overall there is considerable variation. Many AHe ages substantially exceed corresponding AFT ages, and the oldest exceeds the age of the granite. All labs returned data sets with statistically nonoverlapping values, and in contrast to



**Figure 3.** AHe data. Error bars show 95% confidence intervals. (a) S1 ages, with published data from Gautheron et al. (2009). (b) S2 ages, with published data from Mahéo et al., (2013). (c,d) AHe ages versus effective uranium content ( $eU = [U] + 0.235[Th]$ ) for S1 and S2, respectively. (e,f) AHe age versus ESR for S1 and S2, respectively. AHe = apatite (U-Th)/He; ESR = equivalent spherical radius.

Gautheron et al. (2009), no sets of ages pass the  $\chi^2$  test. Some old ages are arguably identifiable as outliers within their data sets, but in others the old ages do not stand out; for L1, for example, the only determinations with overlapping confidence intervals are ~200 Ma. Seven of 26 single-grain age values (27%) are older than 150 Ma, and four of five labs saw at least one age over 150 Ma. A positive age-eU

**Table 5**  
Apatite (*U-Th*)/He Data for Sample S1

Lab	Grain <sup>a</sup>	F <sub>T</sub> <sup>b</sup>	ESR <sup>c</sup> (μm)	N <sub>pyr</sub> <sup>d</sup>	U (ppm)	Th (ppm)	Sm (ppm)	eU (ppm)	He (nmol/g)	Age <sub>corr</sub> (Ma)	± 1 SE (Ma) <sup>e</sup>	Mean age (Ma)	SD age (Ma)	p(χ <sup>2</sup> )
1	1	0.782	65	-	34.4	24.9	355	40.3	57.65	208.9	11.1	157.2	53.5	0
	2	0.790	68.4	-	32.2	28.4	321	38.9	32.16	133.7	7.2			
	3	0.753	58.1	-	34.1	35.6	441	42.5	53.34	192.8	10.4			
	4	0.767	62.8	-	23.3	33.4	423	31.2	14.26	93.3	5.0			
6	1	0.754	65.2	2	16.9	32.1	363	24.4	16.89	124.8	14.7	128.6	39.5	0
	2	0.74	61.4	2	14.1	18.6	294	18.4	20.06	195.4	23.0			
	3	0.742	62.4	1	43.1	42.7	558	53.2	32.89	112.4	13.2			
	4	0.729	58.9	2	35.7	32	382	43.2	28.46	120	14.1			
	5	0.763	67.8	2	19	32.1	329	26.6	13.27	90.6	10.7			
8	1	0.721	55.4	-	41.8	61.2	403	56.2	24.88	106.7	6.4	105.9	10.7	0.02
	2	0.711	53.7	-	21.1	38.5	414	30.1	13.75	106.2	6.4			
	3	0.741	59.3	-	30.6	31.4	367	38	14.94	90.7	5.4			
	4	0.782	71.6	-	35.6	42.2	442	45.6	21.94	105	6.3			
	5	0.723	55.9	-	17.6	28	320	24.2	12.74	121	7.3			
9	1	0.767	74.5	2	22.6	31.5	329	30	11.84	93	5.8	128.8	48.6	0
	2	0.821	92.6	2	16.3	6.4	192	17.8	17.64	216.7	13.4			
	3	0.826	96.5	1	19.3	36.5	394	27.9	12.10	94.5	5.9			
	4	0.785	82.3	2	16.4	30.8	320	23.6	12.11	117.3	7.3			
	5	0.801	87.6	2	9.9	14.4	200	13.3	10.57	177.2	11.0			
	6	0.783	71.4	0	28.5	34.4	333	36.6	12.27	77.7	4.8			
	7	0.786	81.8	2	15.2	17.8	252	19.4	12.93	152.1	9.4			
	8	0.805	90	2	22.5	28.4	317	29.2	13.20	101.7	6.3			
11	1	0.786	67.3	1	26	45.8	277	36.8	13.79	86.7	6.0	156.7	149.3	0
	2	0.684	44.8	2	40.9	46.1	544	51.8	76.26	380.6	27.1			
	3	0.755	58.2	1	35.1	44.1	329	45.4	15.06	79.9	5.6			
	4	0.713	48.8	2	19.6	14.3	358	23	7.25	79.6	5.9			
Gautheron et al. (2009)	1 m	0.68	-	-	20.8	30	-	27.9	9.1	87.7	7.0	89.7	2.3	0.9
	2 m	0.73	-	-	18.2	28.2	-	24.8	9.1	92.2	7.4			
	3 m	0.71	-	-	22.7	32.9	-	30.4	10.5	89.2	7.1			
#										26				
Mean										133				
Median										109.6				
Minimum										77.7				
Maximum										380.6				
St. Dev.										65.5				

Note. ESR = equivalent spherical radius.

<sup>a</sup>Grain or aliquot number; m = a multigrain analysis. <sup>b</sup>Fraction of total helium retained after accounting for long alpha particle stopping distances. <sup>c</sup>Equivalent sphere radius. <sup>d</sup>Number of pyramidal terminations of grain. <sup>e</sup>One standard error of the mean corrected grain age.

correlation is present in the L1 data, but not for any other lab (Figure 3c). Neither are there significant correlations between age and grain size (expressed as the equivalent sphere radius, ESR; Figure 3e).

Sample quality may have been a factor in the high degree of dispersion observed for S1. In their data submission, Lab 11 reported that the S1 grains in their aliquot were of “marginal quality”, and they would normally recommend that they not be analyzed. No other lab groups made such a report, but evaluations of grain quality were not requested in exercise instructions.

For sample S2 (5 labs, 32 grains in total), single-grain ages are less scattered and range from 3.8 to 8.4 Ma (Figure 3b, Table 6), with an overall mean age of 6.4 Ma, similar to that reported in the original study (Mahéo et al., 2013). Lab means range from 5.6 to 7.2 Ma, and all labs report single-grain ages bracketing the published value. There are no obvious outliers, but there is more scatter than reported in the original study, with only one data set (L11) passing the  $\chi^2$  test. The grouped age data show no correlations with eU or ESR (Figures 3d and 3f), although some correlations within lab data sets may be inferred, especially if some grains are selectively omitted. Overall U and Th contents are similar among labs, with the exception that the mean U value reported by L11 is significantly higher than that of the others (Figure 3d).

**Table 6**  
Apatite (*U-Th*)/He Data for Sample S2

Lab	Grain	F <sub>T</sub>	ESR (μm)	N <sub>pyr</sub>	U (ppm)	Th (ppm)	Sm (ppm)	eU (ppm)	He (nmol/g)	Age <sub>corr</sub> (Ma)	±1 SE (Ma)	Mean age (Ma)	SD age (Ma)	p(χ <sup>2</sup> )
1	1	0.784	68	-	76.2	111.4	147.3	102.4	3.79	4.56	0.25	6.40	1.18	0
	2	0.740	55.3	-	75.7	62.9	131.4	90.5	4.71	6.31	0.34			
	3	0.785	68.1	-	72.8	103.5	170.2	97.1	3.61	7.82	0.42			
	4	0.739	55.9	-	66.5	90.8	126.0	87.8	4.18	6.56	0.34			
	5	0.770	62.6	-	70.5	49.3	131.5	82.1	3.88	6.76	0.37			
6	1	0.690	52	1	49.4	70.0	144.0	65.9	1.56	4.36	0.51	6.40	1.36	0
	2	0.672	48.7	2	82.1	115.6	134.1	109.2	4.07	6.90	0.81			
	3	0.675	49.7	1	72.4	106.2	127.7	97.3	3.76	7.14	0.84			
	4	0.645	45.5	2	56.4	104.3	106.8	80.9	3.43	7.84	0.92			
	5	0.785	75.4	0	81.2	90.7	120.0	102.5	3.21	5.79	0.68			
8	1	0.789	72.5	-	52.7	31.2	112.3	60.0	1.79	6.89	0.41	7.18	0.68	0.01
	2	0.754	61.5	-	95.2	52.7	152.5	107.6	3.34	7.55	0.45			
	3	0.762	66.2	-	52.8	90.0	127.3	74.0	1.88	6.09	0.37			
	4	0.715	54	-	54.9	79.9	147.0	73.7	1.98	6.86	0.41			
	5	0.712	52.7	-	48.5	44.6	89.0	58.9	1.63	7.10	0.43			
	6	0.754	63	-	72.2	87.9	114.3	92.8	2.96	7.75	0.46			
	7	0.714	54	-	48.5	73.8	97.8	65.8	1.93	7.51	0.45			
	8	0.779	70.8	-	64.1	75.3	133.7	81.8	2.92	8.36	0.50			
	9	0.685	48.4	-	69.6	100.6	131.3	93.2	2.20	6.32	0.38			
	10	0.718	54.6	-	30.5	42.9	95.1	40.6	1.18	7.36	0.44			
9	1	0.750	61.4	0	45.5	41.9	79.3	55.3	1.29	5.74	0.36	5.62	1.18	0
	2	0.796	74.8	0	60.1	46.2	93.7	71.0	1.62	5.29	0.33			
	3	0.800	80.8	1	74.2	38.3	75.6	83.2	1.37	3.80	0.24			
	4	0.726	60.1	1	74.4	111.6	154.9	100.6	2.94	7.39	0.46			
	5	0.796	85.1	2	83.0	88.0	118.5	103.7	2.79	6.21	0.38			
	6	0.728	63.5	2	55.3	54.2	84.6	68.0	1.43	5.31	0.33			
11	1	0.753	57.5	1	129.4	132.7	133.8	160.6	3.67	5.62	0.39	5.71	0.61	0.10
	2	0.734	52.2	2	131.7	19.3	169.6	136.2	2.92	5.41	0.38			
	3	0.788	68.6	2	107.7	163.2	124.1	146.0	3.58	5.74	0.41			
	4	0.741	55.8	2	68.9	93.9	92.1	90.9	2.50	6.86	0.50			
	5	0.707	48.1	2	85.4	74.7	163.0	103.0	2.19	5.55	0.39			
	6	0.698	46.1	2	121.6	44.8	177.5	132.1	2.53	5.08	0.36			
Mahéo et al. (2013)	1	0.76	59	-	103.1	66.5	-	118.7	3.10	6.4	0.4	6.45	0.10	0.98
	2	0.74	54	-	97.9	68.3	-	114	2.91	6.4	0.4			
	3	0.74	55	-	85	58.2	-	98.6	2.61	6.6	0.4			
	4	0.74	54	-	78	40.6	-	87.5	2.25	6.4	0.4			
#									36					
Mean									6.38					
Median									6.44					
Minimum									3.80					
Maximum									8.36					
St. Dev.									1.11					

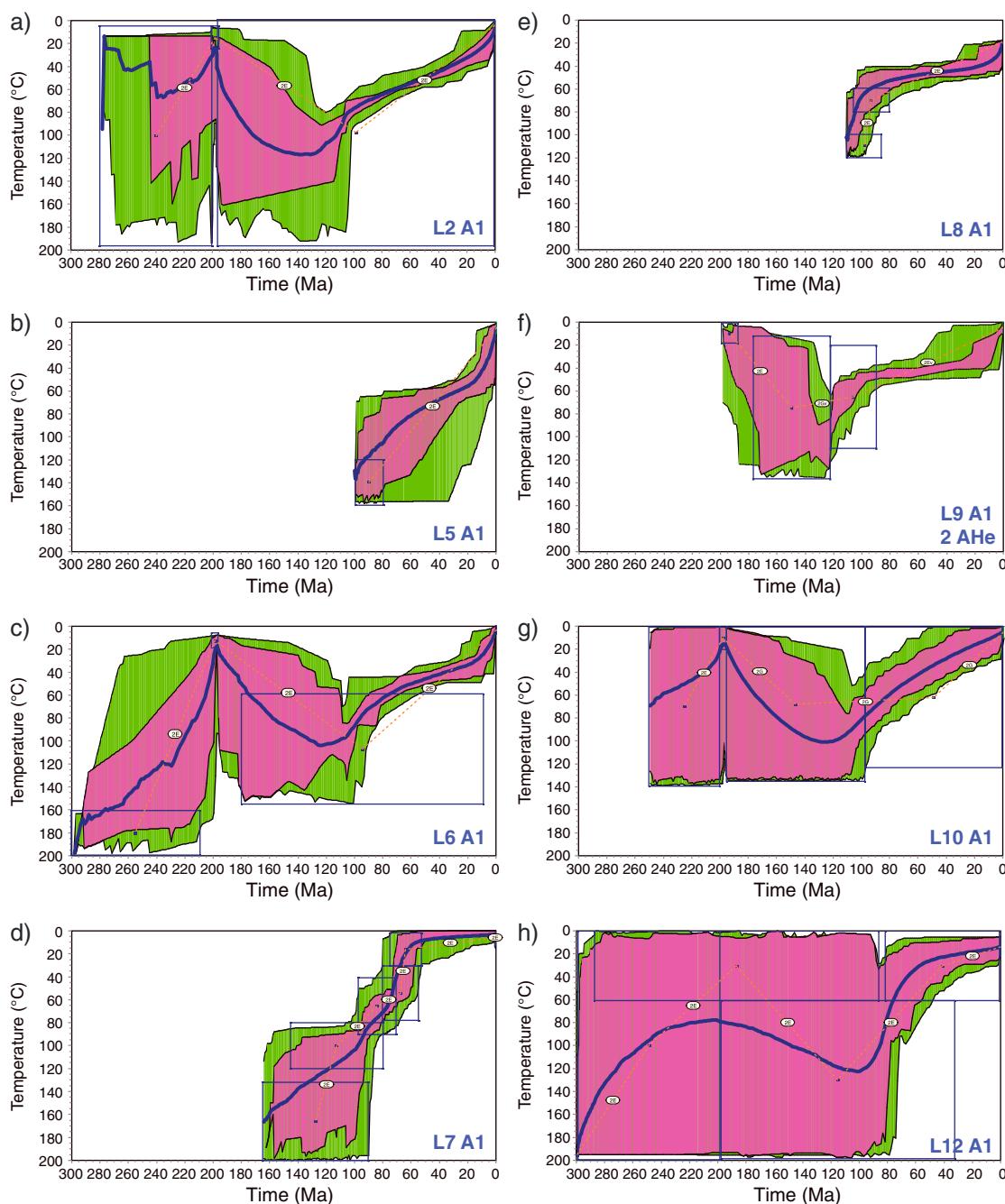
Note. ESR = equivalent spherical radius.

### 3.4. Apatite U/Pb Data

The apatite U/Pb ages reported by L4 were  $302.8 \pm 18.6$  Ma for S1 and  $29.0 \pm 3.9$  Ma for S2. Both are in good agreement with previous age determinations and overlap the crystallization ages of their host granite bodies.

### 3.5. Thermal History Inverse Modeling

Thermal history inversions returned by study participants are shown in Figures 4 and 5 for S1 and Figures 6 and 7 for S2, with time-temperature scaling adjusted and matched to the extent possible to facilitate comparison. HeFTy results (Figures 4 and 6) were mostly submitted as HeFTy-format files containing data and results, allowing model setup and parameters to be examined. QTQt models (Figures 5 and 7) were submitted as static images, with different program outputs for different labs, with text descriptions of how



**Figure 4.** Submitted thermal history inversions for S1 that used HeFTy. Results presented as envelopes surrounding acceptable (green) and good (pink) fitting histories, and dark lines are weighted mean paths averaged at each time. Boxes indicate time-temperature constraints, and connections between boxes are denoted by codes described by Ketcham (2005). AHe = apatite (U-Th)/He.

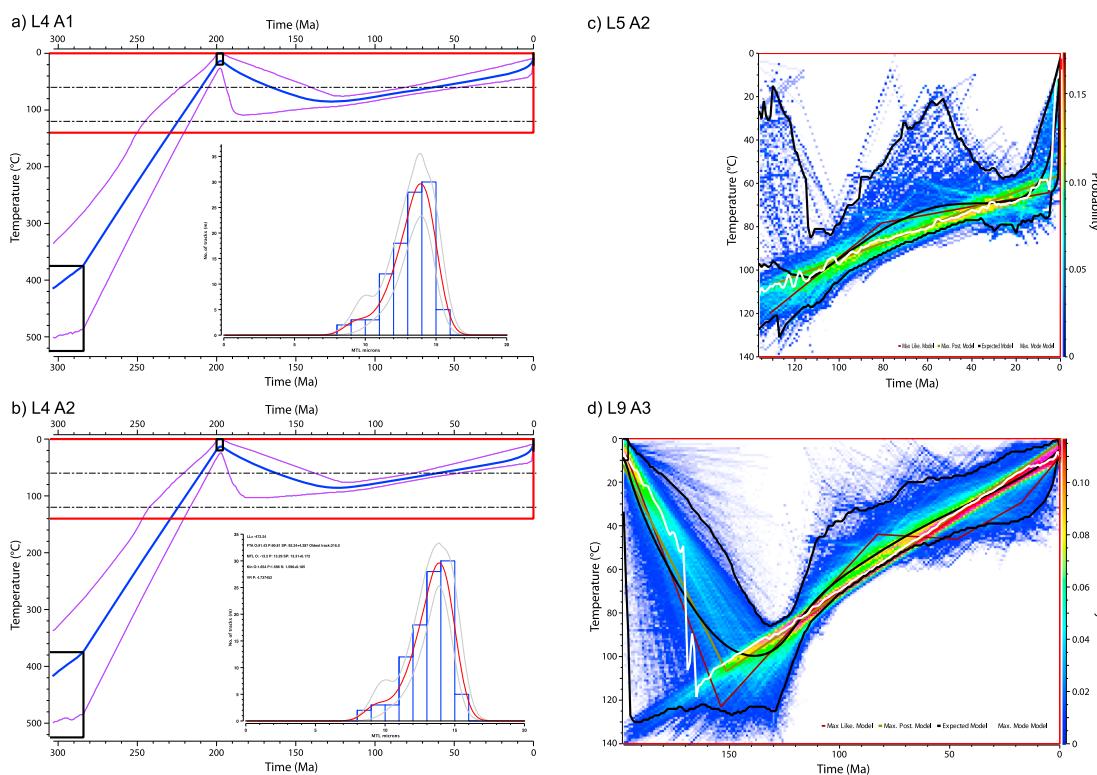
models were run, mostly consisting of where constraints and priors were set and how many burn-in and post-burn-in iterations were used. Table 7 summarizes a number of relevant modeling parameters and outcomes.

## 4. Discussion

### 4.1. Data

#### 4.1.1. AFT Data

The AFT age data are overall quite compatible with each other and the original studies, at least within the large uncertainties typical of fission-track dating, owing to its poor counting statistics compared to mass

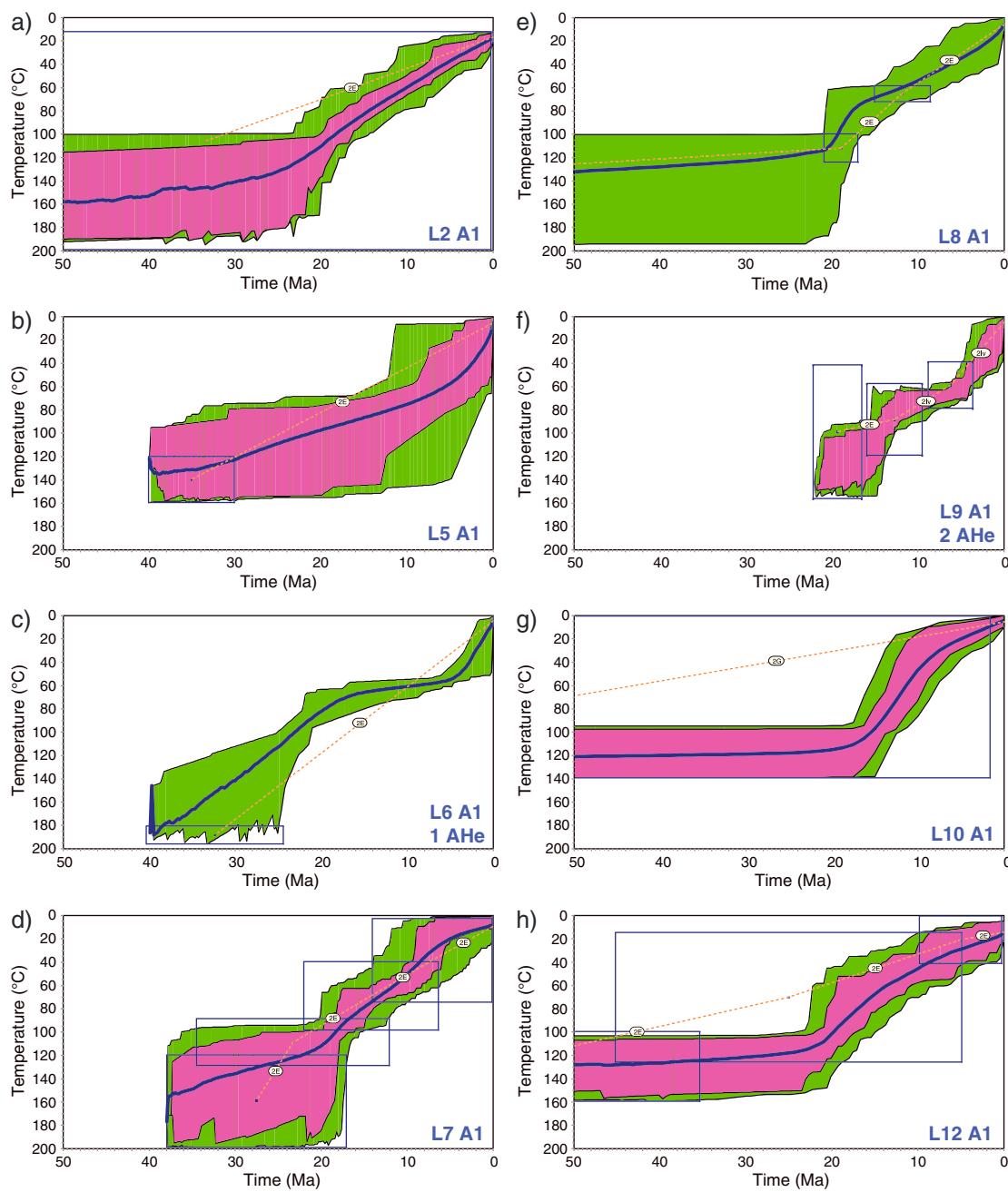


**Figure 5.** Submitted thermal history inversions for S1 that used QTQt. (a) and (b) show Bayesian credible intervals bracketing the expected model, while (c) and (d) also contour the marginal posterior distribution for the thermal history results and show the maximum likelihood and posterior models (Gallagher, 2012).

spectrometry methods. The observed compatibility suggests that uncertainties are scaled appropriately, at least when using the EDM. The correlation of  $\chi^2$  probabilities for the two samples is unexpected, however, and implies that some uncertainties may not be estimated properly. In particular, the LA-ICP-MS method may either have uncertainties that are not well characterized, or there may be implementation details for optimizing LA-ICP-MS analysis of U in apatite that are not adopted uniformly among study participants. The lab group that produced the LA-ICP-MS data with the highest  $p(\chi^2)$  reported multiple laser spots on some grains that reproduced relatively poorly; this averaging may have been responsible for the relatively good performance. At the same time, having both  $\chi^2$  probabilities at 100% (L12) when using EDM analysis is also unexpected, as the expected value for  $p(\chi^2)$  (i.e., the average value over many repetitions) in the model case of samples taken from a uniform population with well-understood uncertainties is 0.5, or 50%.

The observed excess variation in the length data supports earlier recommendations that length measurements need better standardization (Ketcham et al., 2009, 2015). None of the submitted modeling work utilized a lab- or analyst-specific length calibration, although it should be recognized that the recommendation for one (Ketcham et al., 2015) was published only after the exercise had begun. Length standard deviations also varied by up to a factor of ~2 for both samples, and the different length distributions would also be expected to affect thermal history modeling. The ability of c axis projection to increase congruence among analysts in terms of both mean length and standard deviation supports earlier conclusions that accounting for anisotropy is helpful in diminishing inconsistencies (Ketcham et al., 2007a, 2009).

$D_{\text{par}}$  values show even more severe excess variation, including multiple values outside the range used for the underlying annealing models. Carlson et al. (1999) reported  $D_{\text{par}}$  values ranging from 1.59 to 1.89  $\mu\text{m}$  for five F-apatites and Durango apatite, and the minimum  $D_{\text{par}}$  value reported by Barbarand, Carter, et al. (2003) is 1.62  $\mu\text{m}$ . In the present study, four analysts reported mean values below 1.5  $\mu\text{m}$  for S1. Other measurements for S1 ranged up to 2.79  $\mu\text{m}$ , and the range was 1.55 to 3.07  $\mu\text{m}$  for S2. Because  $D_{\text{par}}$  can be used to infer both annealing kinetics and initial length, this scale of variation has tangible impact on inverse modeling, as discussed below.

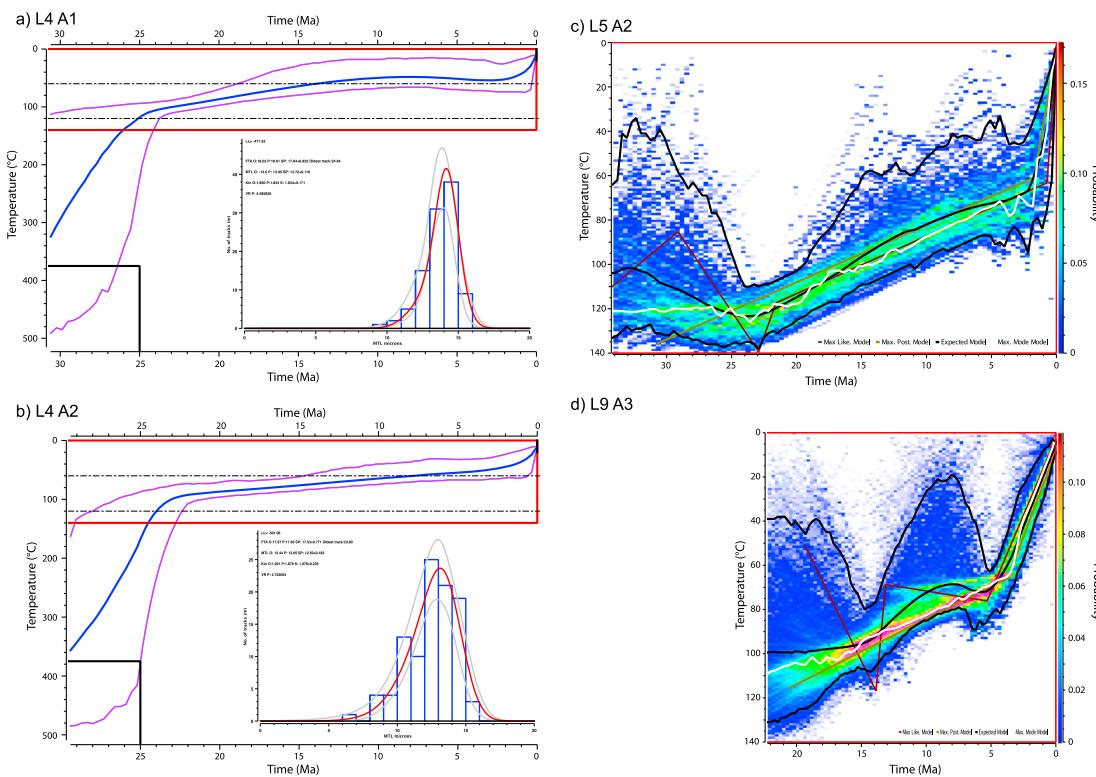


**Figure 6.** Submitted thermal history inversions for S2 that used HeFTy. AHe = apatite (U-Th)/He.

#### 4.1.2. AHe Data

Both samples were selected from among many analyzed in previous studies for their expected relatively minimal AHe age variation. The much greater observed dispersion, from all labs, implies that the lack of dispersion in the original studies may in part have been a statistical effect of small sample sizes. If one takes only a limited number of samples from a dispersed population over and over (i.e., in a many-sample study), once in a while only nondispersed grains will be chosen, making that sample appear deceptively well behaved.

Another likely important distinction is that both prior studies used aliquots of two to four grains per analysis, whereas all participants in the present study provided single-grain ages. This is likely to have been particularly important for S2, as the overall mean across all grains in all labs matched the results from four-grain aliquots, and all labs reported ages dispersed to either side of the overall mean. However, the overall mean or median



**Figure 7.** Submitted thermal history inversions for S2 that used QTQt.

age is not necessarily the *correct* age; the fact that the dispersion exceeded the expectation based on analytical factors and/or repeated analysis of standards means that we really do not know the reason for it.

As mentioned previously, sample quality may have been an issue for S1. Careful grain screening is an important component of AHe analysis; although general guidelines exist (e.g., Farley & Stockli, 2002), the final determination of what constitutes a good grain for analysis is at the discretion of the individual analysts and lab groups. Given the context that this was a community, interlaboratory exercise, participants may have felt obligated to provide data even if they considered sample material to be substandard, as reported by L11.

#### 4.2. Submitted Thermal History Models

##### 4.2.1. S1, HeFTy Models

The models submitted (Figure 4) show a great deal of variation in their setup and outcomes. Placement of constraint boxes was extremely variable. Three of eight models did not include the nonconformity event, with two making the mistake of having the initial constraint box correspond to the measured AFT age. The wide confidence intervals for all times prior to ~100 Ma in all results indicate that the data do not contain much information in that region. Post-100 Ma, all models reflect cooling, but with a range of different characters: roughly linear (L2 and L10), early cooling followed by quiescence (L7 and L12), or cooling featuring a late, rapid stage (L5, L6, L8, and L9). Each contribution is discussed briefly in the following.

The L2 contribution (Figure 4a) is a typical and well-formed HeFTy model, consisting of an initial constraint to allow cooling to begin before the time of the nonconformity; one representing the nonconformity itself, one for post-nonconformity burial, and one for cooling to the present day. The burial constraint, which defines the range of times to be searched for peak reheating, is larger than it needs to be because the large proportion of long confined tracks indicates that peak burial cannot have been too recent, and in order to cause significant resetting, the peak burial temperature must have been >80 °C. However, the only penalty for such a wide constraint is a longer model running time, or fewer solutions found, depending on the ending condition used.

The L5 model (Figure 4b) suffered from a data entry problem, in that the value for  $N_d$  (number of tracks counted in the dosimeter) was not entered correctly and the value used was very small. This in turn

**Table 7**  
Selected Parameters From Final Submitted Thermal History Models

HeFTy models					HeFTy input parameters				HeFTy results	
Code <sup>a</sup>	AFT c ax-p <sup>b</sup>	Kin. par. <sup>c</sup>	AHe model <sup>d</sup>	AHe used <sup>e</sup> S1 + S2	$L_0$ <sup>f</sup>	Length GOF <sup>g</sup>	Ending cond. <sup>h</sup>	Paths <sup>i</sup> S1 + S2	Good <sup>j</sup> S1 + S2	Acc <sup>k</sup> S1 + S2
L2 A1	B03	Dpar	-	-	from Dpar	Kuiper	# tries	10000 10000	7 59	70 161
L5 A1	C99	Dpar	-	-	from Dpar	K-S Test	# tries	10000 10000	82 152	358 378
L6 A1	B03	Dpar	F2000	0 of 5 1 of 5	from Dpar	Kuiper	# tries	10000 10000	13 0	72 45
L7 A1	B03	Dpar	-	-	const 16.3	Kuiper	# good	53543 469902	50 82	7625 8113
L8 A1	C99	Dpar	-	0 of 5 0 of 10	from Dpar	Kuiper	# tries	10000 10000	71 0	445 663
L9 A2	B03	Dpar	F2009	2 of 8 2 of 6	from Dpar	Kuiper	n.r.	n.r.	n.r.	n.r.
L10 A1	B03	Dpar	-	-	from Dpar	Kuiper	# good	88337 1675093	1000 1000	3660 4462
L12 A1	B03	Dpar	-	-	from Dpar	Kuiper	# tries	10000 10000	697 70	379 282
QTQt models					QTQt input parameters					
L4 A1 + 2	C99	Cl wt%	-	-	Total iter. <sup>l</sup>	Burn-in <sup>m</sup>				
L5 A2	C99	Dpar	-	-	200000	n.r.				
L9 A3	B03	Dpar	F2000	2 of 8 2 of 6	10000	5000				
					90000	50000				

Note. AFT = apatite fission-track; AHe = apatite (U-Th)/He.

<sup>a</sup>Lab and analyst that conducted the modeling. <sup>b</sup>C axis projection model used from Ketcham et al. (2007a); B03 = based on Barbarand, Carter, et al. (2003); C99 = based on Carlson et al. (1999). <sup>c</sup>Kinetic parameter used for Ketcham et al. (2007a, 2007ab) AFT annealing model. <sup>d</sup>Diffusion parameters for AHe model; F2000 = Farley (2000); F2009 = Flowers et al. (2009). <sup>e</sup>Number of AHe grains analyzed and used in modeling for each sample. <sup>f</sup>Method for estimating initial track length: from Dpar = based on  $D_{\text{par}}L_0$  relation measured by Carlson et al. (1999) or Barbarand, Carter, et al. (2003); const 16.3 = constant value.

<sup>g</sup>Goodness-of-fit statistic used for lengths. <sup>h</sup>Whether model was run for a constant number of tries or until some number of *good* fits found. <sup>i</sup>Paths modeled in final result for each sample. <sup>j</sup>Number of paths in final model for each sample meeting HeFTy good criterion. <sup>k</sup>Number of paths in final model for each sample meeting HeFTy *acceptable* criterion. <sup>l</sup>Total number of models calculated. <sup>m</sup>Models calculated during burn-in stage.

inflated the calculated uncertainty in the AFT age, resulting in overly wide confidence intervals. This model also shows an effect from the large measured  $D_{\text{par}}$  value of 2.79  $\mu\text{m}$ , which raises the estimated closure temperature to 118 °C (from ~100 °C for a typical F-apatite, both assuming ~10 °C/Ma cooling). The higher modeled resistance to annealing requires the sample to remain at a higher temperature to achieve the observed track shortening, necessitating a rapid final episode of cooling to arrive at the present-day temperature.

The L6 model setup is similar to L2, and the result is correspondingly similar (Figure 4c). The burial constraint is smaller and arbitrarily caps the peak reheating temperature at 160 °C, but any t-T path that exceeded that temperature would not retain any information from before that time. L6 acquired AHe data, but because of considerable spread, with only one of five grains younger than the AFT age, it was omitted from the modeling.

The L7 contribution (Figure 4d) utilizes multiple constraints along the cooling path to try to optimize the solution; they have the appearance of somewhat forcing the solution, but they do not exclude any potential paths that have a good fit to the data according to HeFTy criteria. The results were more affected by an issue with how the program responded to incomplete data.  $D_{\text{par}}$  values were only measured and entered for the length data (mean 2.03  $\mu\text{m}$ ), and the single-grain age data were all given a  $D_{\text{par}}$  of zero, resulting in the software adopting a midpoint value of 1.015  $\mu\text{m}$ . This lowered the estimated closure temperature and initial track length, resulting in artificially low temperatures during the latter part of the inversion. As will be seen later in this paper, correcting this error has a considerable effect on the modeled cooling history. In addition, L7 reported the longest measured track lengths for S1.

The L8 results (Figure 4e) also included the problem of only using measured  $D_{\text{par}}$  for the track-length measurements; moreover, they do not honor the present-day temperature constraint and include constraints intended to represent both AFT and AHe ages. Such constraints are superfluous if the respective data are included, as they provide only a simplified and possibly incorrect version of the information already embedded in the model through inclusion of the data. The model shown only includes AFT data, as a separate attempt to include all of the AHe grains resulted in no successfully fitted paths because the single-grain age data do not agree to within uncertainties.

The L9 model (Figure 4f) is routine and included two AHe grains, again after an earlier attempt to use all grains was unsuccessful. It also used standard He-diffusion kinetics (Farley, 2000) rather than a radiation-damage model (e.g., Flowers et al., 2009; Gautheron et al., 2009). The reported justification for this choice was that no age-eU correlation is apparent in the data. However, such reasoning is not valid: if one takes as settled that radiation damage affects diffusivity, then the Farley (2000) kinetic parameters are only appropriate for an apatite with radiation damage equivalent to Durango apatite, and an apatite with different radiation damage will have different kinetics. Just because radiation damage can affect diffusivity does not mean that measured ages will vary with eU in all cases. A recently exhumed rock that only had a brief time for accumulating radiation damage relative to entry into its helium partial retention zone will still be best represented by diffusivity parameters corresponding to low radiation damage, whereas Durango apatite has relatively high damage due to its ~32 Ma near-surface residence and high Th content (McDowell et al., 2005). At the same time, the effect of radiation-damage accumulation and annealing on helium diffusivity in apatite remains an area of active research (Willett et al., 2017), and it would be a mistake to presume that diffusivities calculated with either the Flowers et al. (2009) or the Gautheron et al. (2009) models are definitive.

The L10 and L12 models both feature large constraint boxes and both found a large number of paths fitting their data, providing a nice illustration of where the data resolve the history and where they do not (i.e., where the solution envelopes fill t-T space as much as allowed by the constraints). Another factor contributing to the wideness of the confidence intervals is that no limitation was placed on the maximum allowable heating or cooling rate. The L12 inversion also suffered from the zero- $D_{\text{par}}$  problem, contributing to the strong convex-down shape of the early stages of its final cooling path, and does not honor the present-day temperature specification.

#### 4.2.2. S1, QTQt Models

The submitted QTQt models (Figure 5) also showed significant variation, in many ways reminiscent of the HeFTy results. Comparison is partly hindered by the different program outputs returned; the contour lines from L4 (Figures 5a and 5b) show the Bayesian credible interval bracketing the expected model, which is obtained by averaging the paths tested and accepted; the corresponding contours and path are shown in black in Figures 5c and 5d.

The two models for L4 (Figures 5a and 5b) are very similar in large part because both utilized the same length measurements. The inversion results suggest that the data fairly tightly constrain the post-nonconformity thermal history, with temperatures that did not far exceed 100 °C and featuring roughly linear cooling after ~125 Ma. However, nothing in the AFT data require such a limit on burial heating, and all other HeFTy and QTQt models that cover this part of the history permit higher temperatures. This outcome appears to result from the emphasis that QTQt places on path simplicity leading to a potentially misinterpretable result: since heating to above 110 °C followed by cooling is a more complex path, the algorithm favored paths that do not include this possibility. In this case, appropriate interpretation may be aided by using the *oldest track* model output to see how far back in time the AFT data really constrain the history.

The model for L5 A2 (Figure 5c) features fast, late cooling from over 60 °C. As with the HeFTy model by L5 A1 (Figure 4b), this outcome is caused by the large  $D_{\text{par}}$  value requiring high annealing temperatures. The option in QTQt of allowing  $D_{\text{par}}$  to vary from its prior (the measured value) to help overall model optimization did not help in this case, and in fact sent the solution in the wrong direction: the QTQt model converged on a  $D_{\text{par}}$  of 2.81 μm, rising from a measured prior value of 2.34 μm.

L9 A3 did QTQt modeling in two rounds: first with all AHe grains measured, and then with a limited number of grains (Figure 5d), after the first models failed to reproduce some of the AHe ages. The criteria chosen for the second round was to use only grains with two pyramidal terminations that had an age

less than the AFT age. The Farley (2000) diffusion kinetics were used due to the lack of an age-eU correlation. The final result suggests higher burial temperatures; the highest probability regions suggest roughly linear cooling, although the credible interval (black lines) suggests rapid initial cooling that subsequently slowed.

We note that in both HeFTy and QTQt models, starting times varied from the AFT age through the age of the nonconformity to a time before that. Only the earliest of these starting times is correct: even though the models do not resolve the t-T history prior to maximum burial, this part of the history can contribute short tracks to the overall track-length distribution, and potentially affects later He-diffusion behavior due to damage accumulation. It should therefore be included in the thermal-history inversions to predict more consistently the well-resolved part of the history.

#### 4.2.3. S2, HeFTy Models

As with S1, there is substantial variation in model setup, but the model results are overall more similar, due in part to their relative simplicity. Most feature roughly continuous cooling after ~20 Ma, although the details of the cooling trends differ in this case as well, with some featuring decelerating cooling rates (L7, L10, and L12) and others accelerating (L6, L9, and perhaps L5).

In the L5 model, an improper  $N_d$  value again led to large confidence intervals, and the large  $D_{par}$  required accelerated late cooling. An attempt by L6 to fit three AHe grains along with the AFT data resulted in no fits, and the final model used a single AHe grain with the AFT data, using Farley (2000) diffusion kinetics. As with S1, the models produced by L7 and L10 suffered from incomplete  $D_{par}$  data. The degree of fit of the L8 model was hampered by an unusual track-length distribution that was truncated on the high side; this is an unfavorable arrangement when using Kuiper's test as the goodness-of-fit metric for lengths, due to its greater sensitivity to the tails of the distribution (Press et al., 1988).

#### 4.2.4. S2, QTQt Models

As with S1, QTQt models also show variation. The L4 results both support rapid initial cooling followed by slower cooling. Although overall similar, with overlapping credible intervals, the expected path of L4 A1 implies >10 Ma of stability near or below 60 °C, while in the model of L4 A2, cooling is more linear. The L5 model again features significant late-stage cooling (~30 °C/Ma) due to the large  $D_{par}$  value, allowing QTQt to vary  $D_{par}$  resulted in insignificant change in this case. L9 again modeled in two rounds, first attempting to incorporate most AHe ages and then omitting all but two. This model also features rapid late cooling (~15 °C/Ma.), mainly due to the relatively short track lengths measured in comparison to the other labs (Figure 2d).

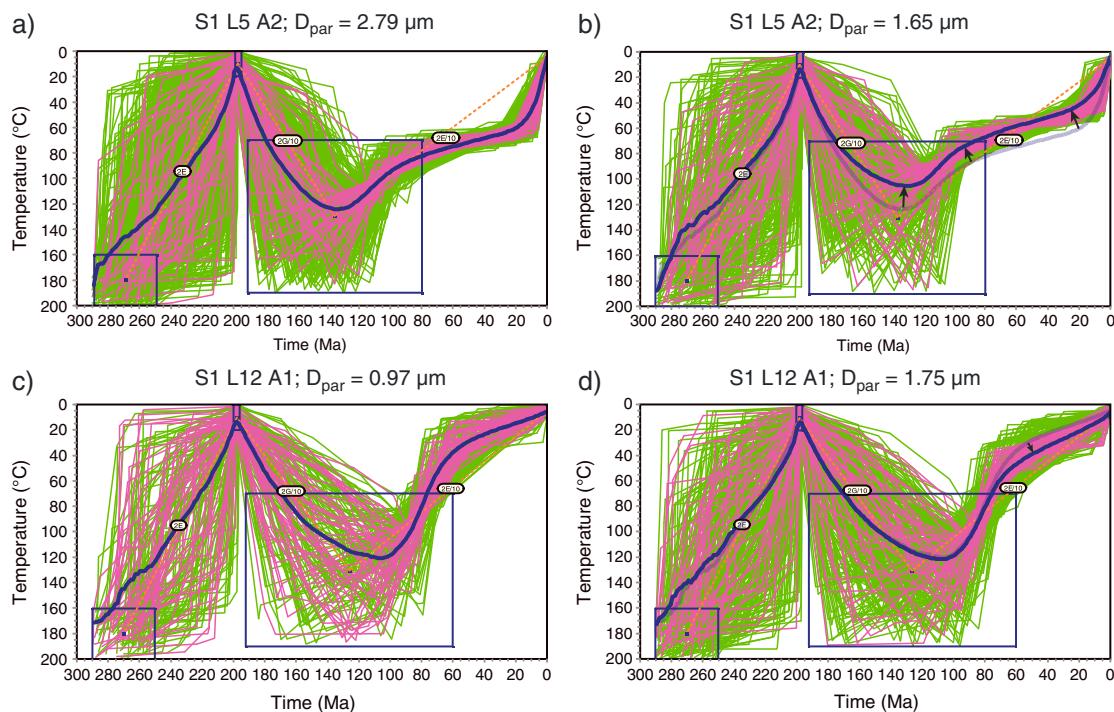
### 4.3. Effects of AFT $D_{par}$ and Length Calibration

Although several labs reported that they had performed length and/or  $D_{par}$  calibrations against natural standards, none of the models submitted reflected the use of such calibrations to adjust model input. As shown here (Figure 2) and previously (Ketcham et al., 2015), length and  $D_{par}$  measurements show considerable variation across the community, and this variation has direct effects on thermal-history reconstruction.

#### 4.3.1. $D_{par}$

$D_{par}$  can impact two things, depending on modeling options chosen: the model annealing kinetics and model initial length ( $L_0$ ). Figures 8a and 8b illustrate the effects of high  $D_{par}$  values, using the L5 A2 model for S1. The high  $D_{par}$  value shifts the post-120 Ma cooling history to temperatures that are ~20 °C higher than in the same model with  $D_{par}$  set to a more typical value for F-apatite. This displacement in turn requires faster and greater late-stage cooling to reach present-day conditions.

At the low- $D_{par}$  end, effects are more subdued because HeFTy treats values below 1.6 μm as equivalent to 1.6 μm for the annealing models, but does allow lower  $D_{par}$  to predict shorter  $L_0$  values to attempt to account for interlaboratory variation. Figures 8c and 8d illustrate the effect of low  $D_{par}$  on the S1 L12 result, in which  $D_{par}$  was inadvertently reduced by including zero values in the input. The erroneous cooling path (Figure 8c) requires early cooling and long residence below ~40 °C. Setting  $D_{par}$  values for dated grains to the average value measured by that analyst with track lengths (Figure 8d) makes cooling more gradual. Although the corrected models in Figures 8b and 8d still diverge somewhat, they are much more similar after accounting for this issue.



**Figure 8.** Effects of changing  $D_{\text{par}}$  on thermal history reconstruction. (a) Inversion from L5 data using the submitted  $D_{\text{par}}$  value, showing good paths in fuchsia, acceptable paths in green, and weighted mean path in dark blue. (b) Inversion from L5 data using a  $D_{\text{par}}$  value more indicative of F-apatite. Semitransparent path and arrows show weighted mean path from (a) and its displacement, showing how change in  $D_{\text{par}}$  affects both high- and low-T history. (c) Inversion from L12 data in which  $D_{\text{par}}$  value is miscalculated because some data were entered with  $D_{\text{par}} = 0$ . (d) Inversion from L12 data in which blank  $D_{\text{par}}$  values are replaced by the mean nonzero value. Semitransparent weighted mean path from (c) and arrow shows effect on low-T part of history.

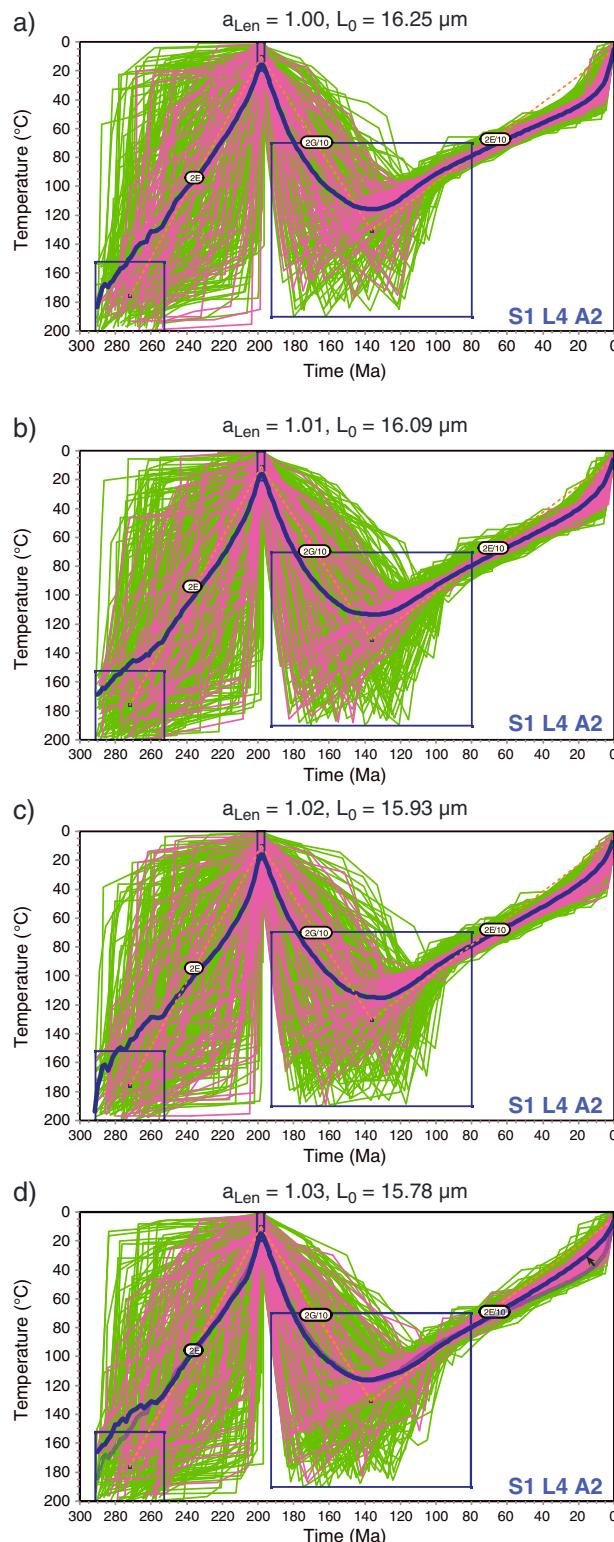
#### 4.3.2. Length Calibration

All of the submitted HeFTy models used a program default for the initial track length, either a constant value of 16.3  $\mu\text{m}$ , based on Durango apatite measurements by Green et al. (1986) and Carlson et al. (1999), or a hardwired calibration of initial track length to  $D_{\text{par}}$  based on either Carlson et al. (1999) or Barbarand, Carter, et al. (2003). Use of such defaults is dangerous; in a previous interlab study, Ketcham et al. (2015) documented significant variation in initial track length reported by 30 lab groups, with a range of 15.3 to 16.8  $\mu\text{m}$  and a mean of  $15.9 \pm 0.1 \mu\text{m}$ , notably different from the previously published values. Ketcham et al. (2015) proposed an adjustment factor,  $a_{\text{Len}}$ , which is the ratio of a published mean track length to the analyst's measured one, to link the values typical of a given analyst to the original measurements underlying the annealing models.

Figure 9 shows the effect of varying  $a_{\text{Len}}$  on the S1 model for L4 A2. The hardwired initial track-length value for the midpoint CI content is 16.25  $\mu\text{m}$ . Increasing  $a_{\text{Len}}$  from 1.00 (no correction) to 1.01, 1.02, and 1.03 roughly places  $L_0$  in the upper, middle, and lower end of the confidence interval of values reported by Ketcham et al. (2015). The effects on the thermal history are small but noticeable, evolving the cooling path from one requiring a final stage of accelerated cooling to one where cooling is almost linear. This change in cooling history occurs almost exclusively in the  $<60^\circ\text{C}$  range, as a longer imposed initial track length requires slightly higher temperatures until late in the thermal history to attain the observed length distribution.

#### 4.4. Effects of Combining AFT and AHe Data

Overall, the effects of combining AFT and AHe data in a single inversion are expected to be varied—in cases where AHe and AFT data are congruent, there may be little change, or the thermal history bounds may be tightened in the temperature ranges where the respective thermochronometers are most sensitive. If AFT and AHe data are incompatible, or AHe data are not internally consistent, then HeFTy would fail to find a fit and return a solution, while QTQt would return a solution that requires careful evaluation and interpretation.



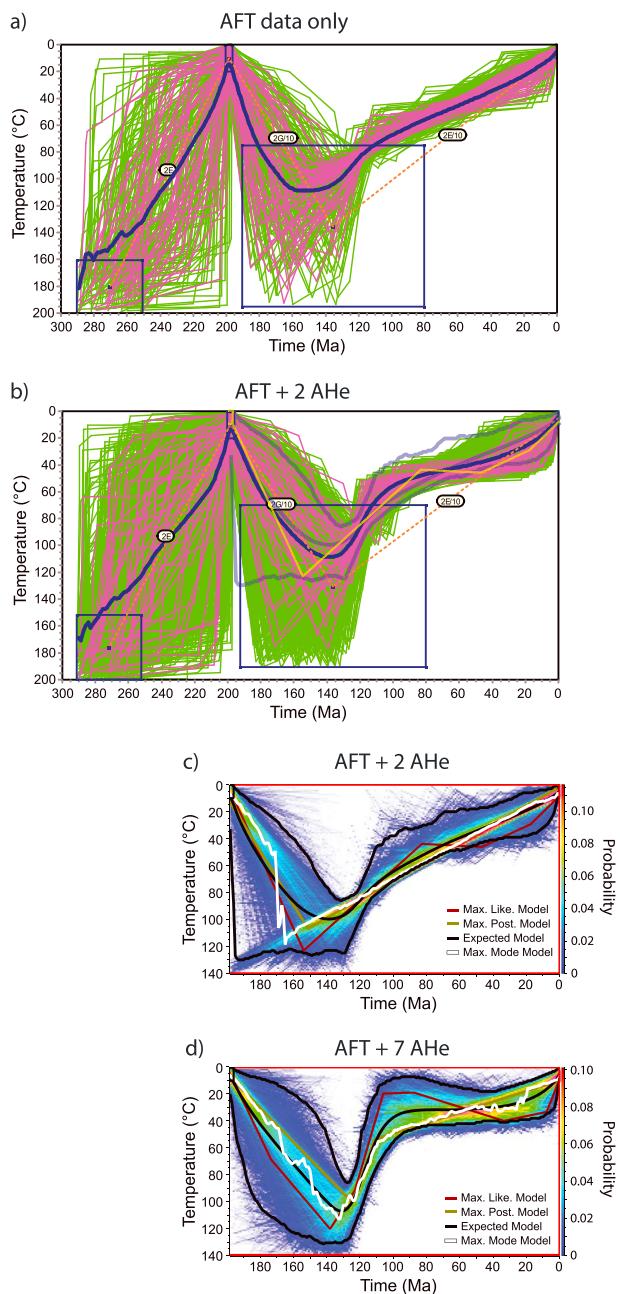
**Figure 9.** Effects of changing assumed value for initial track length ( $L_0$ ) on thermal history reconstruction. The  $a_{\text{Len}}$  is the ratio of the initial track length measured in the data set underlying the annealing model to an analyst-specific value (Ketcham et al., 2015). Semitransparent path in (d) is weighted mean path from (a), with arrow showing net shift in low-temperature history.

Figure 10 shows an instructive example using the L9 contribution. Figures 10a and 10b show HeFTy models generated by RAK, whereas Figures 10c and 10d show the second- and first-round QTQt models generated by L9 A3, respectively. The AFT-only model (Figure 10a) suggests roughly linear cooling following peak burial reheating. Adding in the two AHe grains that have two pyramidal terminations and ages younger than the AFT age (i.e., the second-round model from L9) produces a cooling history that includes an extended interval at a fairly constant temperature, bracketed by two episodes of faster cooling (Figure 10b). The two AHe ages are 101.7 and 93.0 Ma, and thus the effect of including them was to require slightly earlier cooling, which in turn was partially compensated with respect to the AFT data by staying at elevated temperature for a slightly longer period. The corresponding QTQt model (Figure 10c) is quite similar to the HeFTy result; the QTQt 95% credible intervals and expected model are superimposed in blue, and the maximum likelihood model in yellow, on the HeFTy model (Figure 10b) to facilitate comparison. The fitting criteria in the QTQt model prefer a simpler, linear cooling path, which is reflected in the QTQt posterior probabilities and gives a sense akin to the HeFTy model without AHe data (Figure 10a), but the maximum likelihood path and the credible interval bounds are similar to the nonlinear cooling path in Figure 10b. When seven AHe grains are modeled in QTQt (Figure 10d), the cooling history changes markedly. The three of the seven grains that have ages older than the AFT age are not reproduced at all closely (supporting information S1), and one of the younger ages is also not well fitted. The overall cooling path has been distorted to include greater and more rapid initial cooling, probably in an attempt to reduce the misfit with the old outliers.

This example highlights the different meanings of the primary visualizations used to convey HeFTy and QTQt results. HeFTy colors reflect the degree of fit between model and data, and because there is no path consistent with all of the data to within their respective 95% confidence intervals, HeFTy produces a blank diagram if all seven AHe grains are included (not shown). In contrast, the two QTQt models (Figures 10c and 10d) both show similar degrees of maximum probability on the color scale, despite the vastly different degrees of fidelity to the data included. This is because the probability shown in QTQt, formally called the marginal posterior distribution, is that of a given coordinate being contained among those paths tested and accepted during the course of the inversion. It is conditional on which paths are tested, with QTQt favoring simpler ones; which paths are accepted, which is dependent on relative, not absolute likelihood (i.e., the best the algorithm finds, even if the best is not very good); and the relative frequency of testing and accepting different path types. Thus, a high probability is not diagnostic of a good fit, and the match to the individual data elements must be evaluated as well (Gallagher, 2016).

#### 4.5. Effects of Modeling Parameter Selection

The diversity of contributed inverse-model results arises from a combination of differences in the input measurements, the modeling software used and its underlying methodology, and differences among analysts (including occasional errors) in model setup. As a final test of the reproducibility of thermal history inversion, we model data from all labs that contributed AFT-length data using the same model setup in



**Figure 10.** Effects of adding AHe data for thermal history reconstruction. (a) AFT-only model from L9 data. (b) Same model with two AHe grains retained in second-round QTQt modeling by L9 A3 (c); semitransparent blue lines show 95% credible interval and expected model, and yellow line shows maximum likelihood model, from QTQt inversion. (d) First-round QTQt modeling by L9 A3 using seven AHe grains. AFT = apatite fission-track; AHe = apatite (U-Th)/He.

In general, accelerated early post-burial cooling (relative to the overall post-burial cooling path) was required by the data contributed by most labs, with the L8 data predicting the most nonlinear cooling path and the L4 data leading to an essentially linear path. A plurality of results supports roughly linear cooling since ~70 Ma (L2, L6, L7, L9, L10, and L12), while others show significantly faster cooling over the final ~20 Ma. (L5 and L8), with L4 results being intermediate. These differences are strongly influenced by length and  $D_{\text{par}}$  calibration (Figures 8 and 9).

HeFTy, correcting errors in model input but not adjusting measured values (e.g.,  $D_{\text{par}}$ ). AHe grains were included for L8 and L9, but not for L6.

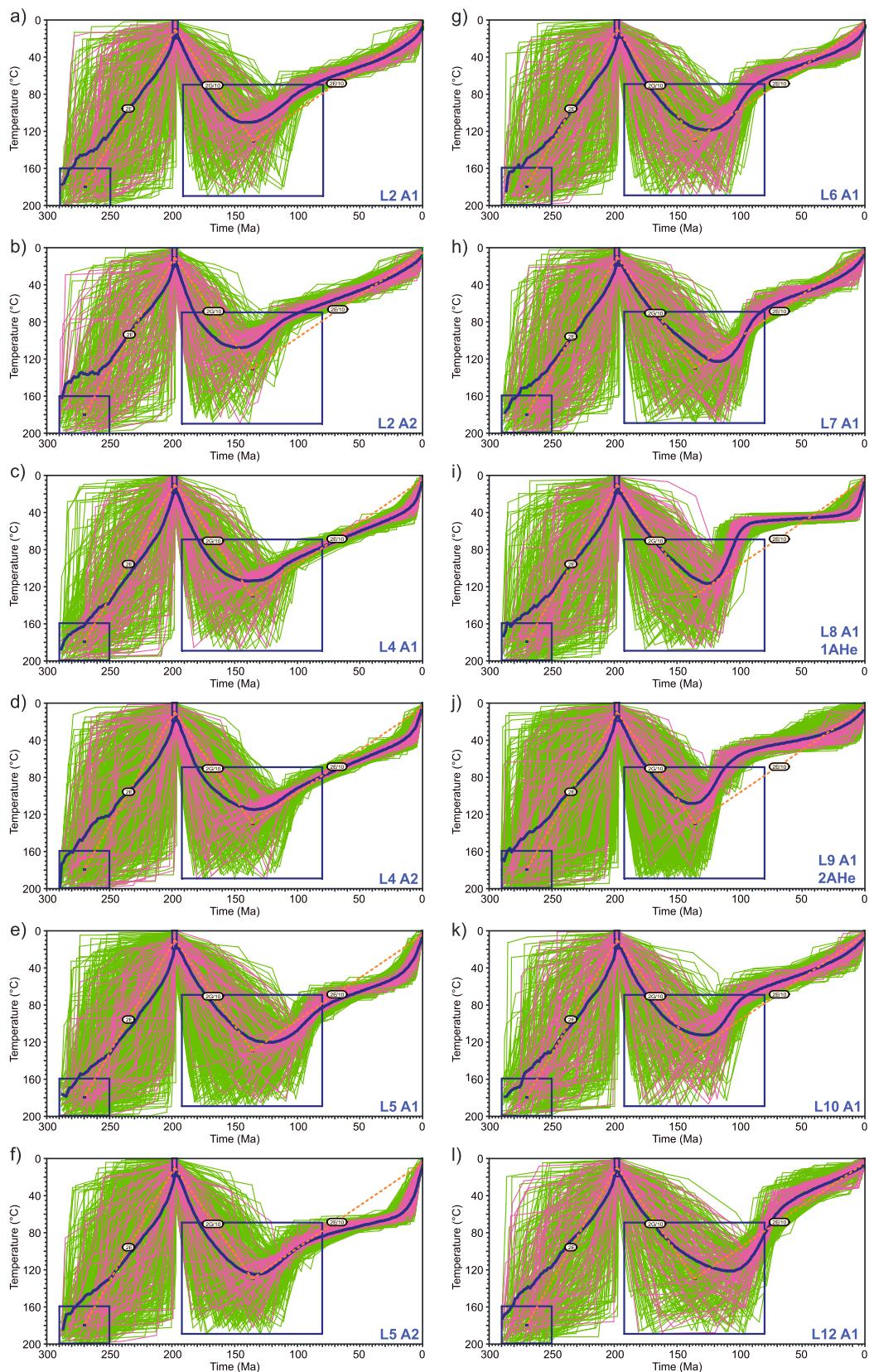
We note in advance that we had several advantages over study participants in this remodeling exercise, including knowing crystallization ages and general tectonic settings and thus expectations for maximum reasonable heating and cooling rates. We purposefully minimized the amount of such information we provided to avoid biasing the analytical portion of this exercise, and participants cannot be faulted for not incorporating them into their modeling efforts. Part of our intention, however, is to demonstrate good modeling practices, especially when using HeFTy, and thus how such additional information is advantageously incorporated.

We further note that L11 opted not to provide an inverse modeling analysis of either of their data sets because they considered the information provided on tectonic context and crystallization ages to be insufficient for making sensible data interpretations. This choice reflects another interesting topic, that of how to best approach inverse modeling. The extent to which inverse models should contain no or minimal independent data to avoid biasing the results, as opposed to incorporating independent data to avoid geologically nonsensical results and better refine model predictions and ground interpretations, remains a topic of active discussion (Gallagher & Ketcham, 2018; Ketcham et al., 2018).

#### 4.5.1. S1 Models

In HeFTy modeling, it is best practice for each constraint to have a specific purpose and to spell out the reasoning (e.g., Flowers et al., 2015). For S1, the following HeFTy inversion setup was used: an initial, high-temperature constraint corresponding to the initial unroofing history, a low-temperature constraint corresponding to the Hettangian nonconformity, a broad constraint to encompass possible times and temperatures for maximum subsequent burial, and a present-day constraint. For the post-nonconformity period, a maximum heating/cooling rate of 10 °C/Ma was imposed to avoid generating histories that are geologically unlikely for the region.

As might be expected, the results (Figure 11) are much more similar than models run by individual analysts (Figures 4 and 5). Convergence of paths indicates that the principal thermal history information concerns the period after the onset of cooling from maximum burial. Inversions can be compared based on the shape of the cooling path, and in particular whether and where cooling accelerated. Early-stage cooling needs to be evaluated carefully, as it is partly influenced by the HeFTy model setup allowing arbitrarily high burial temperatures; thus, the trend is only well defined once the family of solutions becomes closely grouped (e.g., at about 110 Ma in Figure 11a). Also, although it is natural to be drawn to subtle differences in the weighted-mean path curves, the good paths as defined by HeFTy criteria all fit the data to a high degree, so that a range of cooling path types may be consistent with a specific data set, and apparently disparate inversions may in fact feature considerable overlap.



**Figure 11.** HeFTy modeling of data from all contributors that submitted apatite fission-track data for S1, using a consistent model setup. Models utilizing AHe data are denoted by number of AHe grains used. AHe = apatite (U-Th)/He.

Different aspects of the data control the different stages of cooling. The timing of onset of cooling is controlled mostly by the AFT age, and therefore varies with the ages. The strongest controlling factor for the shape of the cooling path is the shape of the AFT length distribution, although, as temperature falls, the impact of calibration rises. Inclusion of AHe data in L9 leads to more punctuated cooling compared to AFT-only models (Figure 10); the same occurs to a lesser extent for L8.

#### 4.5.2. S2 Models

The S2 models (Figure 12) were set up with the benefit of the apatite U/Pb data contributed by L4. The cooling was divided into two stages: an early stage allowing postcrystallization cooling at any rate and a later stage where the maximum cooling rate was limited to 20 °C/Ma. Some data sets, particularly L4 A1, L5 A2, and L9 A1, would allow even faster late cooling. Constraint boxes to divide these two components of the history have identical placement, except for the two labs that reported younger AFT ages (L9 and L10), which required moving the lower time boundary to more recent times to avoid over-constraining the solution. Two results for L8 are included, one for AFT data only and the other including the median L8 AHe age.

Differences among model results are analogous to those seen for S1, and can be interpreted in the same ways, by comparing the timing of onset and the variability in cooling rates. Differences in the timing of onset of cooling are again controlled by the AFT age, which did not correlate with S1 results (Figure 1); in other words, no lab returned a systematically old or young age compared to its peers for both samples. On the other hand, interestingly, many differences in cooling path shape correlate between the two samples for a given lab. Results for labs 2, 6, and 7 are similarly linear throughout their extent in both samples; labs 10 and 12 suggest initially slightly faster cooling that decelerates; and labs 4, 5, and 9 require accelerated late cooling, which for L9 is accentuated by the AHe data included.

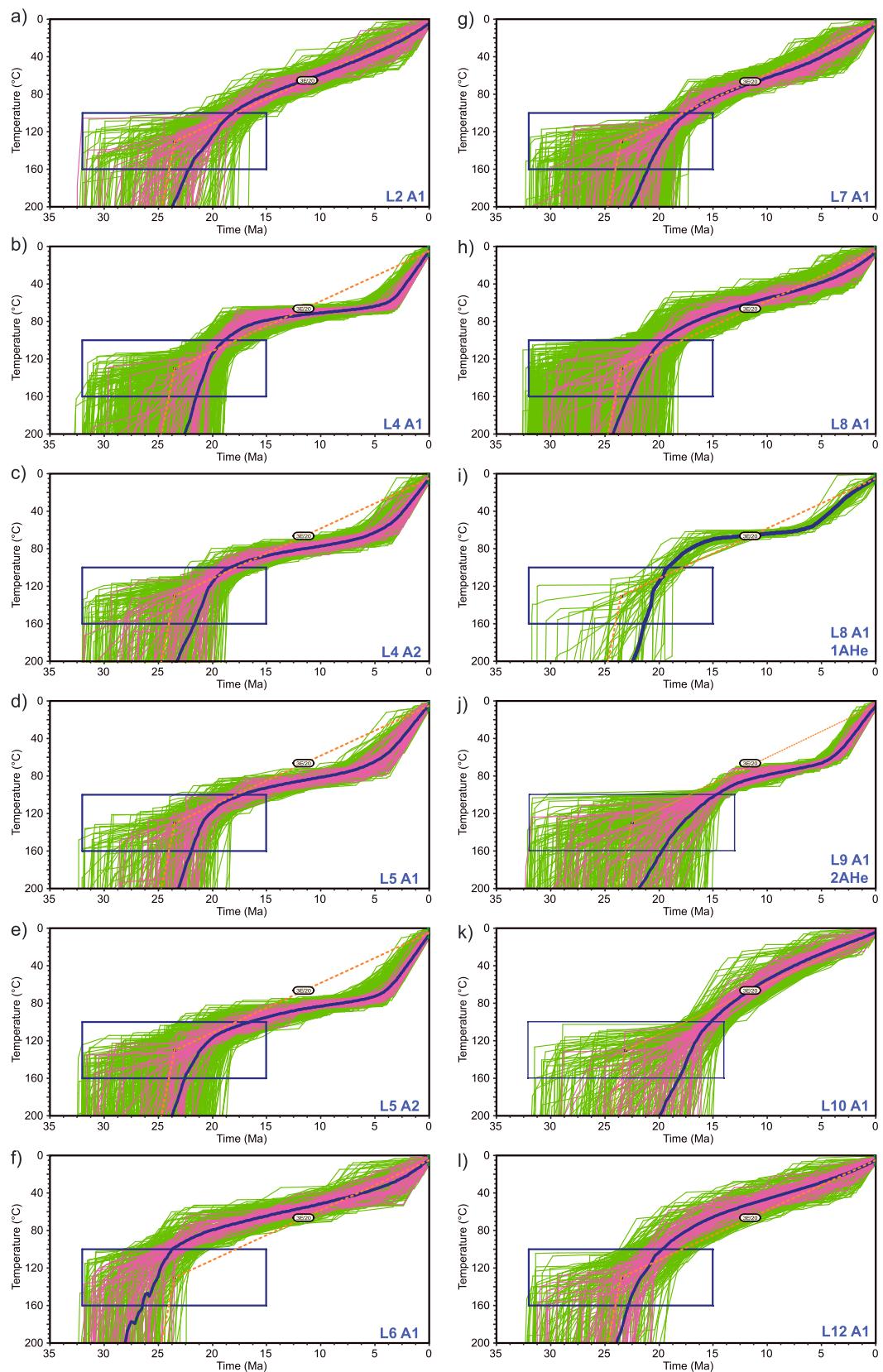
This correlation implies that a significant component of the differences among modeled cooling paths may be due to laboratory-specific practices, certainly including length and  $D_{\text{par}}$  calibration, and possibly including other aspects of length measurement that may influence biasing (relative likelihood of observation and measurement) among different track populations. The correlation of thermal history results is probably in some ways already reflected in the correlation of track lengths (Figure 2e), but not directly; path shape is controlled not only by changes in mean track length, but also in length distribution as well. For example, S2 models for L6 and L12 have similar shapes, but the mean c axis projected length for L6 is 2SE below the S2 median and that for L12 is 2SE above it.

Adding AHe data in the results of L8 and L9 also had analogous effects for both samples: inversions that included AHe data tended to feature more accentuated late and early cooling, with a slower cooling interval in between, compared to AFT-only models. In the original S2 model for L6 (Figure 6c), adding the AHe data also had this effect, compared with the AFT-only inversion in Figure 12f. Including the AHe grain in the new HeFTy L6 model for S2 was unsuccessful; however, because we employed radiation-damage controlled (Flowers et al., 2009) rather than standard (Farley, 2000) AHe diffusion kinetics, the resulting higher closure temperature (caused by the high eU of the sample) changed predicted system behavior such that the AHe and AFT data were no longer compatible. It is not straightforward to assess what caused this misfit, however; L6 returned the oldest S2 AFT age in the study, and if the study-average AFT age of 17.9 Ma is used instead, the AFT and AHe data become compatible.

#### 4.6. HeFTy Versus QTQt Models

Overall, when the input data are self-consistent (i.e., can be fitted within errors), predicted cooling histories are similar between the two programs in this study. As typically used, HeFTy employs more complex t-T paths, not to find better or more apparently precise fits but for the opposite purpose of widening the confidence intervals to defend against overly optimistic interpretation. The proper amount of complexity for a geological thermal history remains an interesting and open question.

The ability of QTQt to allow kinetic parameters to vary acknowledges that our understanding of annealing/diffusion kinetics, and our ability to infer them from our measurements, is imperfect, particularly when using  $D_{\text{par}}$  for AFT. However, in this study this ability did not lead to an improved estimate of kinetics; of the two divergent  $D_{\text{par}}$  analyses from L5, one was left essentially unchanged by QTQt and the other was made to depart even further from the group. This occurred in part because no other data were used in the



**Figure 12.** HeFTy modeling of data from all contributors that submitted apatite fission-track data for S2, using a consistent model setup. Models utilizing AHe data are denoted by number of AHe grains used. AHe = apatite (U-Th)/He.

L5 inversions and the main avenue for improving an estimate of kinetics is to increase consistency with other measurements, from other samples and/or other data types.

Another pertinent difference is in the treatment of AHe data for which the dispersion exceeds the expectation based on analytical factors and reproducibility of standards. If no path fits each piece of the data to within its uncertainties, HeFTy returns no result, whereas QTQt optimizes whatever fit there is. In both cases undertaken by L9, human interpretation and intervention by culling the data was necessary to produce a reasonable result. In S1, the dispersion toward old ages probably reflects some not-yet-understood process, including such ages in a calculation that presumes the system is adequately characterized is not warranted, and led to a predictably poor result (Figure 10d). In the case of S2, where the dispersion across all labs is generally symmetrical about the group mean, the best fit among grains appears reasonable, although even in this case hand-selecting the data to include was necessary (supporting information).

#### 4.7. Characterizing and Modeling AHe Data

The AHe overdispersion observed in this study leads to the question of what to do about it, from the standpoint of both data acquisition and modeling. Some degree of excess AHe age dispersion is very frequently encountered in current practice (Green & Duddy, 2018). There are many possible reasons for this dispersion, including radiation damage (Flowers, 2009; Gerin et al., 2017; Shuster et al., 2006), crystal size (House et al., 2002; Reiners & Farley, 2001), zoning (Ault & Flowers, 2012; Farley et al., 2011; Hourigan et al., 2005), broken grains (Beucher et al., 2013; Brown et al., 2013), implantation from (U-Th)-rich neighboring grains (Gautheron et al., 2012; Kohn et al., 2008), kinetic variations due to composition (Djimbi et al., 2015; Gautheron et al., 2013), or other unaccounted-for factors such as inclusions and microvoids (Zeitler et al., 2017). Only the first few of these effects have quantitative proposed descriptions that allow reasonable generalized and routine incorporation into inverse models. While other factors are potentially quantifiable, the necessary data usually are not acquired (Cl content and zoning) or acquirable (neighboring grains), while others are not yet well understood (microvoids, etc.).

When few grains are analyzed from a naturally dispersed population, the possibility of spurious reproducibility, or even spurious correlations between age and eU or size, is nonnegligible. Such spurious reproducibility may have affected the published studies on our test samples, as discussed above. The group S2 AHe results in Figure 3f illustrate the dangers. If we assume that all analyses are good and the variation is real, then the data from three labs (L1, L8, and L11) show potential multigrain correlations between age and ESR, particularly if a grain or two is omitted. However, the apparent correlations are not mutually compatible, and in the combined data set, there is no support for a significant overall correlation. A thermal history inversion utilizing these correlations can in turn feature a spurious convergence of solutions and appearance of precision. The only reliable means of avoiding such spurious correlations is to analyze more grains per sample, or to determine a means of screening samples or analyses that flags them effectively (e.g., McDannell et al., 2018).

The fact that even the better-behaved AHe sample in this study had dispersion well in excess of lab-reported uncertainties suggests that these uncertainties may be systematically underestimated, perhaps because they are based on reproducibility measured in standards that are chosen because they are unusually well-behaved (e.g., McDowell et al., 2005). If we take the overall group behavior of S2 as a sign that the dispersion is nonbiased (i.e., divergence is as likely to be positive as negative), and we again presume that all reported analyses are good, then we can use the S2 data to estimate a community-scale level of uncertainty. To achieve a  $p(\chi^2)$  of 0.5, it would require an average 1SE uncertainty of 21.3%, and a probability of 0.05 requires 17.5%. These uncertainties are approximately double to triple the frequently used values of 6–8% based on reproducibility of analyses in standards, or the equivalent of an AFT age with  $N_s = 22$  or 33, respectively, assuming all other errors are secondary.

The proper way to include overdispersed AHe data in thermal history inverse modeling is another interesting and contentious topic. For example, Flowers and Kelley (2011) and Green and Duddy (2018) propose that the safest and most straightforward approach is simply to not use such data for modeling at all. The approach shown here, of picking the subset of grains that are consistent with independent (AFT) data, can easily and justifiably be criticized as cherry picking, but it does represent an attempt to include rather than ignore those data and what information they may contain in the interpretation process, in a quantitative fashion. In cases

where the dispersion is symmetrical and modest, as in S2, averaging ages, using the QTQt maximum likelihood approach, or increasing the single-grain uncertainties for HeFTy modeling in a manner similar to the thought experiment in the previous paragraph are possibilities. In all such cases, these steps should be accompanied by specifically reporting and quantitatively evaluating the dispersion, and its effect on the modeling process and results.

All thermal history inversion modeling results are contingent upon their many underlying assumptions, and our knowledge of no thermochronometric system is perfect. Having to take additional measures to include AHe (or any other) data may weaken the results, but does not necessarily invalidate them; in effect, it just lengthens the list of contingencies. Whether the benefit exceeds the cost can be evaluated on a case-by-case basis, but doing so relies critically on transparency and thoroughness in reporting the inverse modeling process (e.g., Flowers et al., 2015).

#### 4.8. Comparison With Previous Studies at Sample Localities

Although the present exercise primarily concerned the reproducibility of thermal history analysis, the correctness of these results is naturally also of interest. The previous work reconstructing thermal histories at both localities consists primarily of thermochronological analyses, and thus there is some circularity in using them to evaluate the outcomes of this study. The primary advantage of the previous work is the context provided by nearby samples at different elevations. The S1 studies (Barbarand et al., 2001; Gautheron et al., 2009) contain enough scatter in their AFT and AHe age-elevation relationships that a range of cooling histories is compatible with them. However, the Mahéo et al. (2013) AHe age-elevation data support a fast-slow-fast cooling sequence for S2. The majority of the inverse models, when performed consistently, also support this sequence (Figure 12). It is notable that, for some labs, the AFT data alone produce this pattern (L4, L9, and L5 even after removing  $D_{\text{par}}$  effects as in Figures 8a and 8b; see supporting information), while for other labs, the addition of AHe is necessary (L6 and L8). In this case, AFT and AHe both contribute valuable information, and they work best when combined.

### 5. Conclusions and Recommendations

The results of this exercise lead us to the following conclusions and recommendations:

- In AFT thermochronology, the use of  $D_{\text{par}}$  without calibration is hazardous to quantitative interpretation. Length calibration also affects modeled cooling paths, particularly the low-temperature ( $<40^{\circ}\text{C}$ ) parts. However, with sufficient intercalibration, AFT-based thermal histories are potentially reliable down to Earth-surface temperatures. Recommended procedures for calibration of length and  $D_{\text{par}}$  measurements have been proposed recently (Ketcham et al., 2015; Sobel & Seward, 2010); this study underlines their importance.
- AFT ages measured using LA-ICP-MS showed signs of greater dispersion than estimated based on analytical factors. Protocols to identify poorly behaving grains and analyze them with a second or third ablation spot may ameliorate this problem.
- Variability in AFT-length-based modeling results were correlated across the two samples examined here, suggesting that they may be partially linked to lab procedures and/or analyst tendencies. If so, they may be resolvable and reproducibility improved through appropriate community-level calibration.
- Appropriate incorporation of AHe data into inverse modeling may depend on sample-specific reasoning, based on observed levels and modes of dispersion, comparison to other data, and so forth. Lack of dispersion, or presence of kinetic correlations, in a limited number of samples from a larger study can result from undersampling rather than truly reflecting better behavior. The observed level of dispersion in the better-behaved sample in this study is about triple the level typically reported for AHe age analyses, and of similar order to AFT single-grain ages. Further measures to reduce dispersion (e.g., better picking, screening for unusual behavior, and multisample aliquots) or to resolve its causes are called for. Alternatively, a greater number of aliquots per sample could be routinely analyzed.
- He-diffusion kinetics are better described by radiation-damage and annealing models than by a constant diffusivity value, and such models should be used in inverse thermal-history modeling. However, our understanding of He diffusion in apatite and other accessory minerals remains incomplete and the currently available models are not the final word.

- Inverse modeling procedures are as important as analytical reproducibility. Understanding the software, and the way it approaches the inverse problem, is critical for proper model setup and interpretation of results, as is transparency and thoroughness in reporting model setup.

### Acknowledgments

We give great thanks to the lab groups and analysts that provided data and modeling for this study. Mineral separation was performed by François Senebier and Francis Coeur at ISTerre Grenoble, supported by CNRS and Université Grenoble Alpes. K. Gallagher provided comments on an earlier version of the manuscript. We also appreciated constructive reviews by R. Flowers and A. Carter that helped to improve the manuscript. AFT single-grain and track-length data, and all model files returned by study participants and created by the authors for this paper, are provided in the online supporting information.

### References

- Ault, A. K., & Flowers, R. M. (2012). Is apatite U-Th zonation information necessary for accurate interpretation of apatite (U-Th)/He thermochronometry data? *Geochimica et Cosmochimica Acta*, 79, 60–78. <https://doi.org/10.1016/j.gca.2011.11.037>
- Barbarand, J., Carter, A., Wood, I., & Hurford, A. J. (2003). Compositional and structural control of fission-track annealing in apatite. *Chemical Geology*, 198(1–2), 107–137. [https://doi.org/10.1016/S0009-2541\(02\)00424-2](https://doi.org/10.1016/S0009-2541(02)00424-2)
- Barbarand, J., Hurford, A. J., & Carter, A. (2003). Variation in apatite fission-track length measurement: Implications for thermal history modelling. *Chemical Geology*, 198, 77–106. [https://doi.org/10.1016/S0009-2541\(02\)00423-0](https://doi.org/10.1016/S0009-2541(02)00423-0)
- Barbarand, J., Lucaleau, F., Pagel, M., & Séranne, M. (2001). Burial and exhumation history of the south-eastern Massif Central (France) constrained by apatite fission-track thermochronology. *Tectonophysics*, 335(3–4), 275–290. [https://doi.org/10.1016/S0040-1951\(01\)00069-5](https://doi.org/10.1016/S0040-1951(01)00069-5)
- Beucher, R., Brown, R. W., Roper, S., Stuart, F. M., & Persano, C. (2013). Natural age dispersion arising from the analysis of broken crystals: Part II. Practical application to apatite (U-Th)/He thermochronometry. *Geochimica et Cosmochimica Acta*, 120, 395–416. <https://doi.org/10.1016/j.gca.2013.05.042>
- Brichau, S., Respaut, J.-P., & Monié, P. (2008). New age constraints on emplacement of the Cévenol granitoids, South French Massif Central. *International Journal of Earth Sciences*, 97(4), 725–738. <https://doi.org/10.1007/s00531-007-0187-x>
- Brown, R. W., Beucher, R., Roper, S., Persano, C., Stuart, F. M., & Fitzgerald, P. G. (2013). Natural age dispersion arising from the analysis of broken crystals. Part I: Theoretical basis and implications for the apatite (U-Th)/He thermochronometer. *Geochimica et Cosmochimica Acta*, 122, 478–497. <https://doi.org/10.1016/j.gca.2013.05.041>
- Carlson, W. D., Donelick, R. A., & Ketcham, R. A. (1999). Variability of apatite fission-track annealing kinetics I: Experimental results. *American Mineralogist*, 84(9), 1213–1223. <https://doi.org/10.2138/am-1999-0901>
- Cathelineau, M., Boiron, M.-C., Fourcade, S., Ruffet, G., Clauer, N., Belcourt, O., et al. (2012). A major Late Jurassic fluid event at the basin/basement unconformity in western France:  $^{40}\text{Ar}/^{39}\text{Ar}$  and K-Ar dating, fluid chemistry, and related geodynamic context. *Chemical Geology*, 322–323, 99–120. <https://doi.org/10.1016/j.chemgeo.2012.06.008>
- Djimbi, D. M., Gautheron, C., Roques, J., Tassan-Got, L., Gerin, C., & Simoni, E. (2015). Impact of apatite chemical composition on (U-Th)/He thermochronometry: An atomistic point of view. *Geochimica et Cosmochimica Acta*, 167, 162–176. <https://doi.org/10.1016/j.gca.2015.06.017>
- Donelick, R. A., Ketcham, R. A., & Carlson, W. D. (1999). Variability of apatite fission-track annealing kinetics II: Crystallographic orientation effects. *American Mineralogist*, 84(9), 1224–1234. <https://doi.org/10.2138/am-1999-0902>
- Farley, K. A. (2000). Helium diffusion from apatite: General behavior as illustrated by Durango fluorapatite. *Journal of Geophysical Research*, 105(B2), 2903–2914. <https://doi.org/10.1029/1999JB900348>
- Farley, K. A., Shuster, D. L., & Ketcham, R. A. (2011). U and Th zonation in apatite observed by laser ablation ICPMS, and implications for the (U-Th)/He system. *Geochimica et Cosmochimica Acta*, 75(16), 4515–4530. <https://doi.org/10.1016/j.gca.2011.05.020>
- Farley, K. A., & Stockli, D. F. (2002). (U-Th)/He Dating of Phosphates: Apatite, Monazite, and Xenotime. In P. W. Reiners & T. A. Ehlers (Eds.), *Reviews in Mineralogy and Geochemistry* (Vol. 48, pp. 559–577). Chantilly, VA: Mineralogical Society of America. <https://doi.org/10.2138/rmg.2002.48.15>
- Flowers, R. M. (2009). Exploiting radiation damage control on apatite (U-Th)/He dates in cratonic regions. *Earth and Planetary Science Letters*, 277(1–2), 148–155. <https://doi.org/10.1016/j.epsl.2008.10.005>
- Flowers, R. M., Farley, K. A., & Ketcham, R. A. (2015). A reporting protocol for thermochronologic modeling illustrated with data from the Grand Canyon. *Earth and Planetary Science Letters*, 432, 425–435. <https://doi.org/10.1016/j.epsl.2015.09.053>
- Flowers, R. M., & Kelley, S. A. (2011). Interpreting data dispersion and “inverted” dates in apatite (U-Th)/He and fission-track datasets: An example from the US midcontinent. *Geochimica et Cosmochimica Acta*, 75(18), 5169–5186. <https://doi.org/10.1016/j.gca.2011.06.016>
- Flowers, R. M., Ketcham, R. A., Shuster, D. L., & Farley, K. A. (2009). Apatite (U-Th)/He thermochronometry using a radiation damage accumulation and annealing model. *Geochimica et Cosmochimica Acta*, 73(8), 2347–2365. <https://doi.org/10.1016/j.gca.2009.01.015>
- Galbraith, R. F., & Laslett, G. M. (1985). Some remarks on statistical estimation in fission-track dating. *Nuclear Tracks and Radiation Measurements*, 10(3), 361–363. [https://doi.org/10.1016/0735-245X\(85\)90125-5](https://doi.org/10.1016/0735-245X(85)90125-5)
- Gallagher, K. (1995). Evolving temperature histories from apatite fission-track data. *Earth and Planetary Science Letters*, 136(3–4), 421–435. [https://doi.org/10.1016/0012-821X\(95\)00197-K](https://doi.org/10.1016/0012-821X(95)00197-K)
- Gallagher, K. (2012). Transdimensional inverse thermal history modeling for quantitative thermochronology. *Journal of Geophysical Research*, 117(B2), B02408. <https://doi.org/10.1029/2011JB008825>
- Gallagher, K. (2016). Comment on ‘A reporting protocol for thermochronologic modeling illustrated with data from the Grand Canyon’ by Flowers, Farley and Ketcham. *Earth and Planetary Science Letters*, 441, 211–212. <https://doi.org/10.1016/j.epsl.2016.02.021>
- Gallagher, K., Brown, R. W., & Johnson, C. (1998). Fission track analysis and its applications to geological problems. *Annual Review of Earth and Planetary Sciences*, 26, 519–572. <https://doi.org/10.1146/annurev.earth.26.1.519>
- Gallagher, K., & Ketcham, R. A. (2018). Comment on “Thermal history modelling: HeFTy vs. QTQt” by Vermeesch and Tian, *Earth-Science Reviews* (2014), 139, 279–290. *Earth Science Reviews*, 176, 387–394. <https://doi.org/10.1016/j.earscirev.2017.11.001>
- Gautheron, C., Barbarand, J., Ketcham, R. A., Tassan-Got, L., Van Der Beek, P. A., Pagel, M., et al. (2013). Chemical influence on  $\alpha$ -recoil damage annealing in apatite: Implications for (U-Th)/He dating. *Chemical Geology*, 351, 257–267. <https://doi.org/10.1016/j.chemgeo.2013.05.027>
- Gautheron, C., Tassan-Got, L., Barbarand, J., & Pagel, M. (2009). Effect of alpha-damage annealing on apatite (U-Th)/He thermochronology. *Chemical Geology*, 266(3–4), 157–170. <https://doi.org/10.1016/j.chemgeo.2009.06.001>
- Gautheron, C., Tassan-Got, L., Ketcham, R. A., & Dobson, K. J. (2012). Accounting for long alpha-particle stopping distances in (U-Th-Sm)/He geochronology: 3D modeling of diffusion, zoning, implantation, and abrasion. *Geochimica et Cosmochimica Acta*, 96, 44–56. <https://doi.org/10.1016/j.gca.2012.08.016>
- Gerin, C., Gautheron, C., Oliviero, E., Bachelet, C., Djimbi, D. M., Seydoux-Guillaume, A.-M., et al. (2017). Influence of vacancy damage on He diffusion in apatite, investigated at atomic to mineralogical scales. *Geochimica et Cosmochimica Acta*, 197, 87–103. <https://doi.org/10.1016/j.gca.2016.10.018>
- Green, P. F., & Duddy, I. R. (2018). Apatite (U-Th-Sm)/He thermochronology on the wrong side of the tracks. *Chemical Geology*, 488, 21–33. <https://doi.org/10.1016/j.chemgeo.2018.04.028>

- Green, P. F., Duddy, I. R., Gleadow, A. J. W., Tingate, P. R., & Laslett, G. M. (1986). Thermal annealing of fission tracks in apatite 1. A qualitative description. *Chemical Geology (Isotope Geoscience Section)*, 59, 237–253. [https://doi.org/10.1016/0168-9622\(86\)90074-6](https://doi.org/10.1016/0168-9622(86)90074-6)
- Hourigan, J. K., Reiners, P. W., & Brandon, M. T. (2005). U-Th zonation-dependent alpha ejection in (U-Th)/He chronometry. *Geochimica et Cosmochimica Acta*, 69(13), 3349–3365. <https://doi.org/10.1016/j.gca.2005.01.024>
- House, M. A., Kohn, B. L., Farley, K. A., & Raza, A. (2002). Evaluating thermal history models for the Otway Basin, southeastern Australia, using (U-Th)/He and fission-track data from borehole apatites. *Tectonophysics*, 349(1-4), 277–295. [https://doi.org/10.1016/S0040-1951\(02\)00057-4](https://doi.org/10.1016/S0040-1951(02)00057-4)
- Ketcham, R. A. (2005). Forward and inverse modeling of low-temperature thermochronometry data. In P. W. Reiners & T. A. Ehlers (Eds.), *Low-temperature thermochronology* (Vol. 58, pp. 275–314). Chantilly, VA: Mineralogical Society of America.
- Ketcham, R. A., Bernet, M., & Van Der Beek, P. A. (2016). Testing the reproducibility of thermal history analysis. Paper presented at the 15th International Conference on Thermochronology Maresias, Brazil.
- Ketcham, R. A., Carter, A., & Hurford, A. J. (2015). Inter-laboratory comparison of fission track confined length and etch figure measurements in apatite. *American Mineralogist*, 100(7), 1452–1468. <https://doi.org/10.2138/am-2015-5167>
- Ketcham, R. A., Carter, A. C., Donelick, R. A., Barbarand, J., & Hurford, A. J. (2007a). Improved measurement of fission-track annealing in apatite using c-axis projection. *American Mineralogist*, 92(5-6), 789–798. <https://doi.org/10.2138/am.2007.2280>
- Ketcham, R. A., Carter, A. C., Donelick, R. A., Barbarand, J., & Hurford, A. J. (2007b). Improved modeling of fission-track annealing in apatite. *American Mineralogist*, 92(5-6), 799–810. <https://doi.org/10.2138/am.2007.2281>
- Ketcham, R. A., Donelick, R. A., Balestrieri, M. L., & Zattin, M. (2009). Reproducibility of apatite fission-track length data and thermal history reconstruction. *Earth and Planetary Science Letters*, 284(3-4), 504–515. <https://doi.org/10.1016/j.epsl.2009.05.015>
- Ketcham, R. A., Mora, A., & Parra, M. (2018). Deciphering exhumation and burial history with multi-sample down-well thermochronometric inverse modelling. *Basin Research*, 30(Suppl. 1), 48–64. <https://doi.org/10.1111/bre.12207>
- Kohn, B. L., Spiegel, C., Phillips, D., & Gleadow, A. J. W. (2008). Rubbing out apatite helium-age spread in fast-cooled rocks. Paper presented at the 11th International Conference on Thermochronometry, Anchorage, Alaska.
- Lisker, F., Ventura, B., & Glasmacher, U. A. (2009). Apatite thermochronology in modern geology. In F. Lisker, B. Ventura, & U. A. Glasmacher (Eds.), *Thermochronological Methods: From Palaeotemperature Constraints to Landscape Evolution Models* (Vol. 324, pp. 1–23). Geological Society, London, Special Publications. <https://doi.org/10.1144/SP324.1>
- Mahéo, G., Gautheron, C., Leloup, P.-H., Fox, M., Tassan-Got, L., & Douville, E. (2013). Neogene exhumation history of the Bergell massif (southeast central Alps). *Terra Nova*, 25(2), 110–118. <https://doi.org/10.1111/ter.12013>
- Malusa, M. G., & Fitzgerald, P. G. (Eds.). (2018). *Fission-Track Thermochronology and its Application to Geology*. Springer International Publishing AG. [https://doi.org/10.1007/978-3-319-89421-8\\_3](https://doi.org/10.1007/978-3-319-89421-8_3)
- McDannell, K. T., Zeitler, P. K., Janes, D. G., Idleman, B. D., & Fayon, A. K. (2018). Screening apatites for (U-Th)/He thermochronometry via continuous ramped heating: He age components and implications for age dispersion. *Geochimica et Cosmochimica Acta*, 223, 90–106. <https://doi.org/10.1016/j.gca.2017.11.031>
- McDowell, F. W., McIntosh, W. C., & Farley, K. A. (2005). A precise  $^{40}\text{Ar}$ - $^{39}\text{Ar}$  reference age for Durango apatite (U-Th)/He and fission-track dating standard. *Chemical Geology*, 214(3-4), 249–263. <https://doi.org/10.1016/j.chemgeo.2004.10.002>
- Miller, D. S., Crowley, K. D., Dokka, R. K., Galbraith, R. F., Kowallis, B. J., & Naeser, C. W. (1993). Results of interlaboratory comparison of fission track ages for 1992 fission track workshop. *Nuclear Tracks and Radiation Measurements*, 21(4), 565–573. [https://doi.org/10.1016/1359-0189\(93\)90197-H](https://doi.org/10.1016/1359-0189(93)90197-H)
- Miller, D. S., Duddy, I. R., Green, P. F., Hurford, A. J., & Naeser, C. W. (1985). Results of interlaboratory comparison of fission-track age standards: Fission-track workshop—1984. *Nuclear Tracks and Radiation Measurements*, 10(3), 383–391. [https://doi.org/10.1016/0735-245X\(85\)90127-9](https://doi.org/10.1016/0735-245X(85)90127-9)
- Miller, D. S., Eby, N., McCorkell, R., Rosenberg, P. E., & Suzuki, M. (1990). Results of interlaboratory comparison of fission track ages for the 1988 fission track workshop. *Nuclear Tracks and Radiation Measurements*, 17(3), 237–245. [https://doi.org/10.1016/1359-0189\(90\)90041-U](https://doi.org/10.1016/1359-0189(90)90041-U)
- Naeser, C. W., Zimmerman, R. A., & Cebula, G. T. (1981). Fission-track dating of apatite and zircon: An interlaboratory comparison. *Nuclear Tracks and Radiation Measurements*, 5, 56–72.
- Oberli, F., Meier, M., Berger, A., Rosenberg, C. L., & Gieré, R. (2004). U-Th-Pb and  $^{230}\text{Th}$ - $^{238}\text{U}$  disequilibrium isotope systematics: Precise accessory mineral chronology and melt evolution tracing in the alpine Bergell intrusion. *Geochimica et Cosmochimica Acta*, 68(11), 2543–2560. <https://doi.org/10.1016/j.gca.2003.10.017>
- Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1988). *Numerical recipes in C*. Cambridge: Cambridge Univ. Press.
- Reiners, P. W., & Brandon, M. T. (2006). Using thermochronology to understand orogenic erosion. *Annual Review of Earth and Planetary Sciences*, 34, 419–466. <https://doi.org/10.1146/annurev.earth.34.031405.125202>
- Reiners, P. W., & Farley, K. A. (2001). Influence of crystal size on (U-Th)/He thermochronology: An example from the Bighorn Mountains, Wyoming. *Earth and Planetary Science Letters*, 188(3-4), 413–420. [https://doi.org/10.1016/S0012-821X\(01\)00341-7](https://doi.org/10.1016/S0012-821X(01)00341-7)
- Shuster, D. L., Flowers, R. M., & Farley, K. A. (2006). The influence of natural radiation damage on helium diffusion kinetics in apatite. *Earth and Planetary Science Letters*, 249(3-4), 148–161. <https://doi.org/10.1016/j.epsl.2006.07.028>
- Sobel, E. R., & Seward, D. (2010). Influence of etching conditions on apatite fission-track etch pit diameter. *Chemical Geology*, 271(1-2), 59–69. <https://doi.org/10.1016/j.chemgeo.2009.12.012>
- Vermeesch, P. (2017). Statistics for LA-ICP-MS based fission track dating. *Chemical Geology*, 456, 19–27. <https://doi.org/10.1016/j.chemgeo.2017.03.002>
- Wagner, G. A., Miller, D. S., & Jäger, E. (1979). Fission track ages on apatite of Bergell rocks from central Alps and Bergell boulders in Oligocene sediments. *Earth and Planetary Science Letters*, 45(2), 355–360. [https://doi.org/10.1016/0012-821X\(79\)90136-5](https://doi.org/10.1016/0012-821X(79)90136-5)
- Willett, C. D., Fox, M., & Shuster, D. L. (2017). A helium-based model for the effects of radiation damage annealing on helium diffusion kinetics in apatite. *Earth and Planetary Science Letters*, 477, 195–204. <https://doi.org/10.1016/j.epsl.2017.07.047>
- Zeitler, P. K., Enkelmann, E., Thomas, J. B., Watson, E. B., Ancuta, L. D., & Idleman, B. D. (2017). Solubility and trapping of helium in apatite. *Geochimica et Cosmochimica Acta*, 209, 1–8. <https://doi.org/10.1016/j.gca.2017.03.041>