Journal of Archaeological Science: Reports 13 (2017) 609–616

Contents lists available at ScienceDirect

Journal of Archaeological Science: Reports

journal homepage: www.elsevier.com/locate/jasrep

Best practices for calibrating and reporting stable isotope measurements in archaeology

Paul Szpak[a,*], Jessica Z. Metcalfe[b], Rebecca A. Macdonald[b]

[a] Department of Anthropology, Trent University, 1600 West Bank Drive, Peterborough, ON K9L 0G2, Canada
[b] Department of Anthropology, The University of British Columbia, Vancouver, BC V6T 1Z1, Canada

ARTICLE INFO

Keywords:
Isotopes
δ-Values
Calibration
Standards
Precision
Accuracy
Analytical uncertainty

ABSTRACT

The use of isotopic measurements in archaeological research has increased rapidly over the past ~25 years, owing largely to the proliferation of the instruments required to produce these measurements relatively quickly and cheaply. Unfortunately, the understanding of how to adequately calibrate and report these isotopic data has not kept pace. We surveyed nearly 500 archaeological research papers published within the past 25 years that presented original isotopic data. We found that, generally, the majority of studies do not provide adequate information regarding how isotopic measurements were calibrated, nor how analytical uncertainty (precision and accuracy) was assessed. We review and present recommendations for data analysis, calibration, and reporting to aid archaeological researchers who use isotopic measurements and practices. We present a simple method for quantifying standard analytical uncertainty using data that would be provided by most laboratories.

1. Introduction

Isotope ratio mass spectrometry (IRMS) is used widely in archaeological[1] studies to address a variety of questions. Beginning in the late 1990s, the direct interfacing of rapid and automated combustion techniques (e.g., elemental analyzers connected via continuous-flow to IRMS systems) for analyzing bulk organic materials decreased analytical costs and dramatically increased the number of analyses that could feasibly be performed in a given study. Prior to that time relatively few isotopic studies had been conducted in archaeology, and each study produced at most a few dozen measurements. In recent years, an abundance of studies has been conducted, producing thousands of measurements (Fig. 1). Given the now widespread availability of technology to produce isotopic measurements quickly and cheaply, it is important to examine how these measurements are being reported. This is particularly important in archaeology as the researchers primarily responsible for disseminating the results in publications are often not directly involved in obtaining the raw measurements and transforming them into calibrated δ-(delta) values. Moreover, results obtained from commercial laboratories may lack the relevant details or be difficult to interpret with respect to analytical uncertainty, particularly for scholars with a limited understanding of isotope ratio mass spectrometry. A decade ago, Jardine and Cunjak (2005) commented on the increase in laboratories providing isotopic measurements and recognized the potential of a widening knowledge gap between IRMS operators and ecologists disseminating these data. We have noticed a similarly widening knowledge gap in archaeology, particularly as it relates to the reporting of analytical methods and uncertainty. While a number of studies have attempted to examine within- and among-laboratory variation in isotopic measurements, the emphasis has been on sample preparation specifically (e.g., Guiry et al., 2016; Jørkov et al., 2007; Sealy et al., 2014), or more generally on measurements produced by different laboratories (e.g., Pestle et al., 2014). Little attention has been paid to the effects of data calibration or the quantification of measurement accuracy, precision, and overall uncertainty.

The purpose of this paper was fourfold. First, we sought to evaluate the reporting of stable carbon and nitrogen isotopic measurements and their associated uncertainties in the archaeological literature. To do this, we performed a review of relevant literature, focusing on data reporting, calibration methods and quality control (accuracy and precision). The results of this survey suggested that a review of methods and strategies for reporting isotopic data would be useful to archaeologists utilizing IRMS in their research. As such, the second purpose of the paper was to review data reporting and quality control methods and present them in a manner accessible to researchers who are reporting isotopic measurements but not generating the measurements themselves. Third, on the basis of our literature survey and review of

* Corresponding author.
E-mail address: paulszpak@trentu.ca (P. Szpak).
[1] For the purposes of this discussion, we use the terms 'archaeological' and 'archaeology' as catch-alls for isotopic studies in both archaeology and physical anthropology.

http://dx.doi.org/10.1016/j.jasrep.2017.05.007
Received 31 October 2016; Received in revised form 24 April 2017; Accepted 1 May 2017

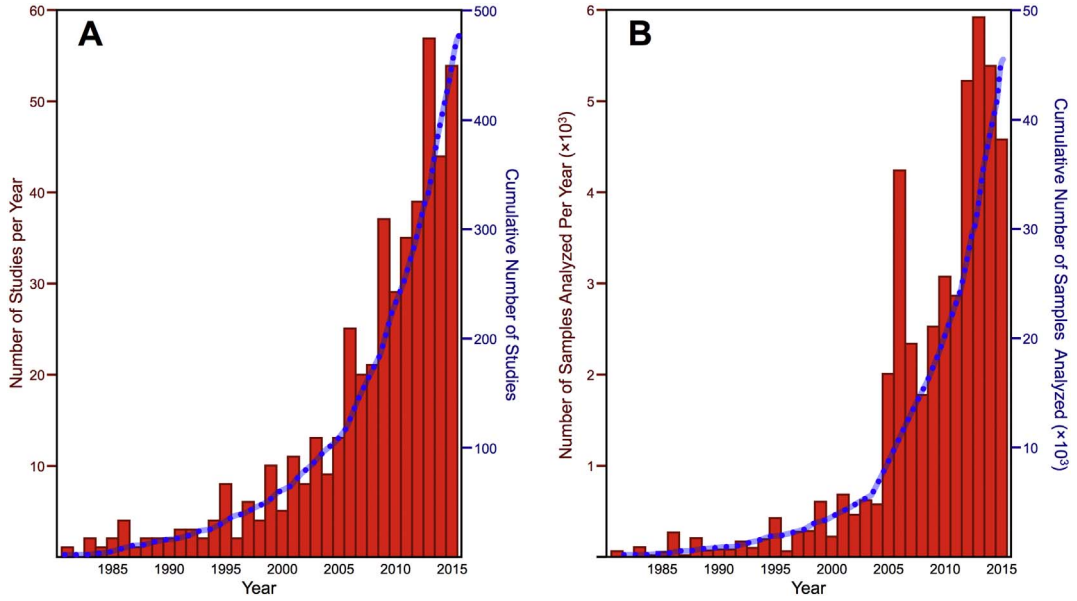

Fig. 1. (A) Number of studies included in the survey per year (primary axis) and cumulative number of studies (secondary axis). (B) Number of isotopic measurements (in thousands) presented in the papers included in the survey per year (primary axis) and cumulatively (secondary axis).

concepts, as well as previously-published International Union of Pure and Applied Chemistry (IUPAC) guidelines (Coplen, 2011), we make a series of recommendations for reporting isotopic data in archaeology. Finally, on the basis of international documents outlining the quantification of standard measurement uncertainty (Joint Committee for Guides in Metrology, 2008; Magnusson et al., 2012), we present a method for determining analytical uncertainty that can be easily derived using a simple set of equations in an Excel spreadsheet. To illustrate examples related to calibration and analytical uncertainty, we have included an example IRMS dataset (Appendix A) that is referenced throughout the paper.

2. Review of key concepts

2.1. Calibration (also referred to as normalization)

Isotopic δ-values are not absolute abundance measurements. Rather, they are relative differences between a sample and an internationally agreed-upon standard (Eq. 1)[2]:

$$\delta = \frac{R_{sample} - R_{standard}}{R_{standard}} \quad (1)$$

where R = the ratio of the heavy to light isotope (e.g., $^{15}N/^{14}N$). The standard mentioned in Eq. (1) is a real or hypothetical international reference material with an accepted value that defines the scale of isotopic measurement for each element. Originally, this standard was PDB (PeeDee Belemnite, a calcareous fossil) for carbon, air N_2 (AIR) for nitrogen, CDT (Canyon Diablo Troilite) for sulfur, and SMOW (Standard Mean Ocean Water) for hydrogen and oxygen (Coplen, 2011; Coplen et al., 2006; Mariotti, 1983). The original samples of PDB, CDT and SMOW have now been exhausted, so the carbon, sulfur, hydrogen and oxygen scales are now reported to VPDB (Vienna PeeDee Belemnite), VCDT (Vienna Canyon Diablo Troilite) and VSMOW (Vienna Standard Mean Ocean Water), respectively (Coplen, 1994; Coplen, 2011; Coplen et al., 2006).

EA (Elemental Analyzer) IRMS systems (or other automated 'online' systems) generate raw values using a single-point calibration relative to a laboratory working gas; the isotopic composition of the working gas is arbitrary (Paul et al., 2007). Even if the 'true' δ value for the reference gas is inputted into the EA software, the isotopic composition of the gas can change over time (Paul et al., 2007). To properly calibrate these raw measurements to internationally-accepted δ-scales, standard reference materials (SRMs) with known isotopic values (previously calibrated to VPDB, AIR, or VSMOW) must be interspersed among samples in each analytical session (or 'run') and then used to calculate a two-point calibration curve (Carter and Fry, 2013; Werner and Brand, 2001). The use of a two-point curve is crucial; by anchoring the raw isotopic values with calibration standards at both the high- and low-ends of the range, δ-values for unknown samples can be shifted and stretched to fit onto the international δ-scales. The laboratory working gas therefore does not need to be calibrated as it only provides an arbitrary comparator for the sample isotope ratios. The calibration of the isotopic measurements occurs entirely after a given analytical session is complete and the raw measured isotopic compositions of the standards can be compared to their known δ values. An example of a two-point calibration curve generated using USGS40 and USGS41 is presented in Appendix A for two different analytical sessions.

Internationally-certified SRMs are preferred for calibration, and can be obtained from organizations such as the IAEA (International Atomic Energy Agency), NIST/NBS (National Institute of Standards and Technology, formerly the National Bureau of Standards), and USGS (United States Geological Survey). These standards have previously been calibrated to the appropriate isotopic measurement scale, and have internationally accepted values assigned to them. Calibration standards should have isotopic values that bracket the high and low ends of the measurement range (Paul et al., 2007). For example, the internationally-accepted SRMs USGS40 and USGS41 are amino acids with $\delta^{13}C$ values of -26.39 and $+37.63$ ‰ and $\delta^{15}N$ values of -4.5 and $+47.6$ ‰, respectively; these values are near or beyond the high and low end of the range of $\delta^{13}C$ and $\delta^{15}N$ values expected for the vast majority of plant and animal tissues (Qi et al., 2003). Internal or in-house SRMs (i.e., standards developed locally and not internationally certified) are less desirable as calibration standards, but are very useful as check standards (see below). If internal SRMs are used as calibration standards, it is necessary to specify their accepted values, and how these values were obtained (Coleman and Meier-Augenstein, 2014). Guidelines for developing in-house standards can be found in Carter

[2] Note that Eq. (1) is not multiplied by 10^3 as per Note 9 in Coplen (2011).

and Fry (2013) and Werner and Brand (2001). In-house and calibration standards should be composed of materials that have similar chemical compositions to the samples being analyzed (referred to as matrix-matching). The importance of matrix-matching varies depending on the isotopes being measured and appears to be less critical for bulk analysis of $\delta^{13}C$ and $\delta^{15}N$ by EA/IRMS systems than for other applications, such as $\delta^{2}H$ measurements of hair (Skrzypek, 2013).

2.2. Quality assurance (analytical uncertainty)

2.2.1. Accuracy

Analytical accuracy is associated with systematic errors in measurement. In other words, how close is a measurement to the 'true' value of the analyte? The accuracy of isotopic measurements can only be evaluated by including SRMs (check standards) in each run that are not used for calibration. Check standards should be calibrated using the identical method applied to the samples. That is, unlike the calibration standards, whose δ-values are treated as known (for the purposes of calibration), the check standards' δ-values are treated as unknowns. The deviation of the measured δ-values from the known values of the check standards demonstrates the degree to which measurements obtained in any given analytical session differ from the 'true' values. These differences represent any systematic errors in the measurements that were not corrected for through calibration and generally should be very small in magnitude.

One potential source of error is instrumental drift, which occurs when systematic errors increase or decrease in magnitude (consistently in the same 'direction', positively or negatively) over the course of an analytical session. The phenomenon is most commonly caused by variations in temperature (Prosser, 1993). Instrumental drift of δ-values within an analytical session can be identified and corrected by analyzing internal SRMs at regular intervals throughout the session to ensure that there is no systematic increase or decrease (Carter and Fry, 2013). Corrections should only be made if the drift is continuous and occurs across the entire analytical session; drift corrections made on only portions of the session can introduce further systematic errors (Ohlsson and Wallmark, 1999; Prosser, 1993). Additional discussion of drift is provided by Merritt and Hayes (1994).

2.2.2. Precision

Analytical precision is generally equated with the repeatability of measurements and random (rather than systematic) errors. In other words, how well can a measurement be reproduced, irrespective of its 'true' value? Precision can be evaluated using SRMs with or without assigned δ-values (e.g., calibration or check standards), and/or with repeated measurement of sample aliquots. While calibration standards cannot be used to monitor accuracy because their average values are set to the known values, the variation of those values can be used as a marker of precision. Precision is frequently measured as standard deviations (generally reported ± 1σ) on SRM values or as absolute differences between repeated sample measurements. The most realistic estimates of precision are obtained using materials that are compositionally similar to the samples being studied. For example, in archaeological studies of bone collagen an internal collagen standard (with ca. 42 wt% C and 15 wt% N) or repeated measurements of collagen samples would provide better estimates of precision than would repeated measurements of a substance such as NBS-22 (a mineral oil with 86 wt% C and no N), even though the latter is an internationally-recognized standard for carbon isotope measurements.

3. Survey of published papers

3.1. Survey methodology

For the purposes of assessing the manner in which analytical uncertainty is discussed in the archaeological literature, a survey of papers presenting original bulk carbon and nitrogen isotopic compositions (the most commonly-used isotopic measurements in archaeology) was conducted. The survey included the journals listed in Table 1. All papers with a publication date between 1980 and 2015 were examined. Several papers were in press at the time of the survey and will likely have a 2016 publication date; consequently, papers with a publication year of 2015 are overrepresented. All papers that included original carbon or nitrogen isotopic data (i.e., new data, not summarized from another source) were included regardless of the nature of the material examined (modern or archaeological). The survey consisted of a series of questions used to determine what details were presented for calibration, analytical accuracy, and analytical precision, and to develop a scoring system to quantify how well they were presented (Appendix B). The scoring system was intended to reflect both the quality and completeness of the details provided. Papers that provided greater detail could receive higher scores than papers that provided less detail, even if the former reported poorer levels of precision or accuracy. The logic behind this approach was that the quality of the results can be better assessed as greater detail on the analytical conditions are provided. Scores were given for three categories: Calibration (0–6), Precision (0–7), and Accuracy (0–4). For calibration, scores of 1–3 were considered an insufficient level of detail, while 4–6 were considered sufficient. For precision, scores of 1–4 were considered insufficient, while 5–7 were considered sufficient. For accuracy, scores of 1–3 were considered insufficient, while a score of 4 was considered sufficient. The details of these scoring systems are provided in Appendix C. Data were also collected on sample type, sample size, instrumentation, and collagen extraction protocol (where applicable). For some papers an answer of 'yes' or 'no' could not be scored for certain categories. For example, it was not possible to determine definitively whether or not all data were presented when the link provided to the Supporting Online Material (SOM) was not functioning. In instances such as this, the result was excluded from the summary statistics for that particular category.

3.2. Survey results

3.2.1. General information

A total of 487 papers were surveyed. A list of every paper included in the literature survey is presented in Appendix D. The majority of the studies reported archaeological data (89%) and most studies analyzed bone collagen (90%). When methodological details on the collagen extraction were provided, 27% of studies used an ultrafiltration step to remove low molecular weight contaminants. Isotopic measurements

Table 1
Journals surveyed to collect data about the presentation of stable isotope data in archaeology and anthropology.

Journal	*n* papers
American Antiquity	20
American Journal of Physical Anthropology	100
Antiquity	8
Archaeological and Anthropological Sciences	11
Archaeometry	5
Current Anthropology	21
Environmental Archaeology	8
International Journal of Osteoarchaeology	36
Journal of Anthropological Archaeology	25
Journal of Archaeological Science	196
Journal of Archaeological Science: Reports	10
Journal of Human Evolution	13
Journal of Island and Coastal Archaeology	10
Latin American Antiquity	6
Oxford Journal of Archaeology	8
World Archaeology	10
Total	487

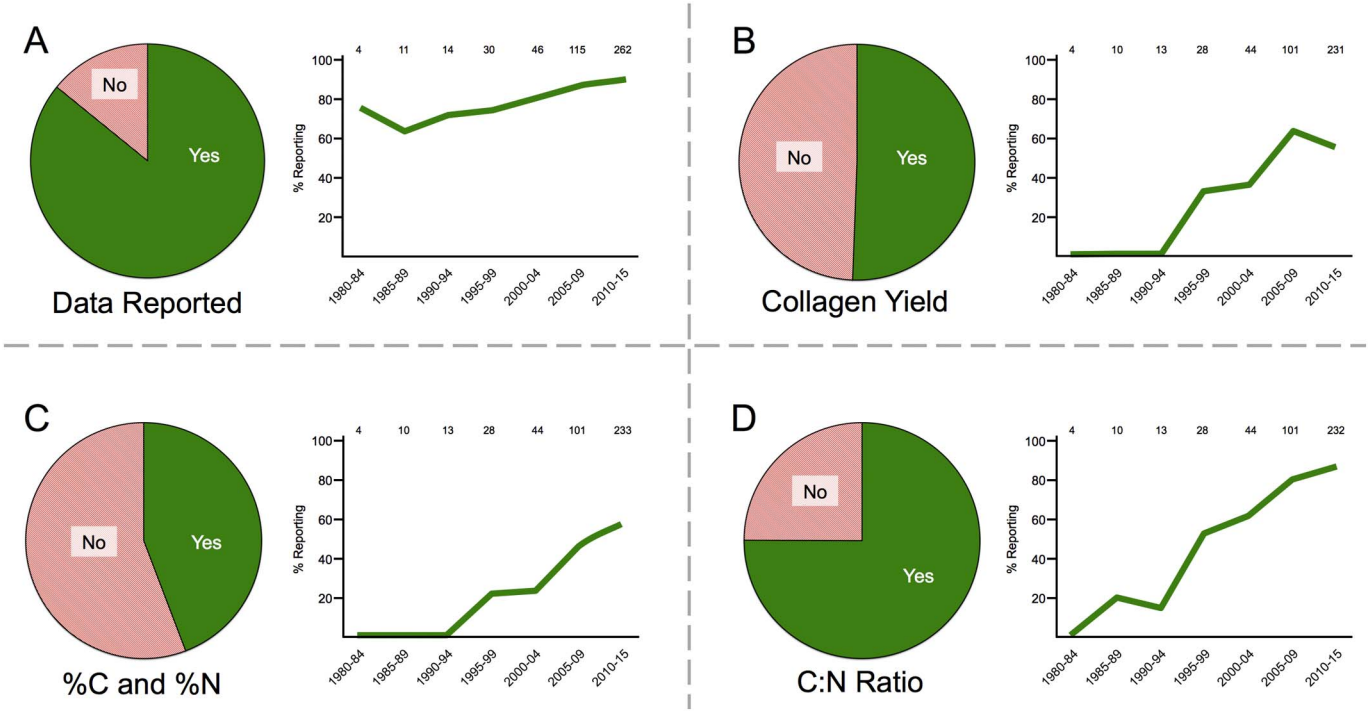

Fig. 2. The percentage of studies reporting: (A) all isotopic data, (B) collagen yields, (C) %C and %N, and (D) C:N ratio. The line graphs show the percentage of studies reporting each parameter at five-year intervals. Note that the final interval represents six years (2010–2015). The numbers above each five-year interval represent the total number of studies.

were generally presented in full with 86% of papers reporting all of their data in the main text or Supplementary material. There was a trend over time towards more complete data reporting (Fig. 2A). In 65% of the studies the laboratory that produced the isotopic data was specifically named in the text (this does not take into account instances in which the laboratory may have been mentioned in the acknowledgments) and 78% of studies named the instrumentation used to produce the measurements.

3.2.2. Bone chemistry and related information

For those studies presenting bone collagen stable carbon and nitrogen isotopic compositions, 79% reported C:N ratios, 45% reported wt% C and wt% N, and 51% reported collagen yields. For each of these three parameters there was a trend towards more complete reporting over time (Fig. 2), which is almost certainly the product of several papers explicitly outlining the importance of disclosing such data (Ambrose, 1990; DeNiro, 1985; van Klinken, 1999).

3.2.3. Calibration

Few studies provided adequate information to determine the method by which raw values were converted to δ-values relative to AIR or VPDB (Fig. 3A). No information regarding calibration was provided in 83% of studies, inadequate information was provided in 10% of studies, and adequate information was provided in only 7% of studies. With respect to calibration standards, 17% of studies mentioned them and 14% specifically named the calibration standards used. For those studies that named their calibration standards, 55% (8% of all studies) used international SRMs certified for calibration (e.g. USGS40, IAEA-CH-6, IAEA-N-1). When in-house SRMs were used for calibration, 21% of those studies provided the known δ-values of those standards. In 7% of studies it was apparent that a two-point calibration curve was used and in 6% of studies it appeared that only a single SRM was used for calibration. It is important to note that the official recommendation from the IAEA to use a two-point calibration was not published until 2006 (Coplen et al., 2006), but the fact remains that in 86% of studies it was not possible to determine whether a one- or two-point calibration curve was used. There was a slight trend towards more complete reporting in recent years with 4% of studies for the period 2005–2009 reporting sufficient information and 10% of studies reporting sufficient information for the period 2010–2015. Nevertheless, it is clear that archaeological studies pay almost no attention to reporting the method of calibration of isotopic results. While it is entirely possible that the vast majority of studies have calibrated their results in accordance with accepted best practices (Carter and Fry, 2013; Paul et al., 2007), the fact remains that this is not possible to determine for 90% of the studies published between 2010 and 2015, when reporting on calibration was at its best.

3.2.4. Accuracy

Of the three areas surveyed, accuracy was by far the most underreported. In 95% of studies accuracy was not mentioned (Fig. 3C). Only 3.5% of the studies surveyed provided adequate information on accuracy, while 1.5% provided some level of information that was deemed inadequate. The lack of reporting regarding analytical accuracy is especially worrying considering the lack of attention apparently paid to calibration, as discussed in more detail below.

3.2.5. Precision

In the papers surveyed, precision was differentiated from accuracy only 4% of the time. Consequently, it was usually impossible to assess whether or not the value that was reported was a specific marker of *precision* or some broader catch-all marker of analytical error. This section includes discussion of these more vague cases as well as more specific references to precision. Because precision, unlike accuracy, can be tracked using calibration standards, check standards, and sample replicates, it can be assumed that precision was monitored by one or more of these markers when they are mentioned in the text.

In 20% of studies, analytical precision (or analytical error more generally) was not mentioned. The level of detail provided for analytical precision was classified as adequate for 22% of studies and inadequate for 58% of studies (Fig. 3B). In 5% of the papers a vague reference was made to long-term precision or error and in 4% of the papers a specific reference was made to long-term or typical laboratory precision or error, sometimes with a non-specific reference to one or

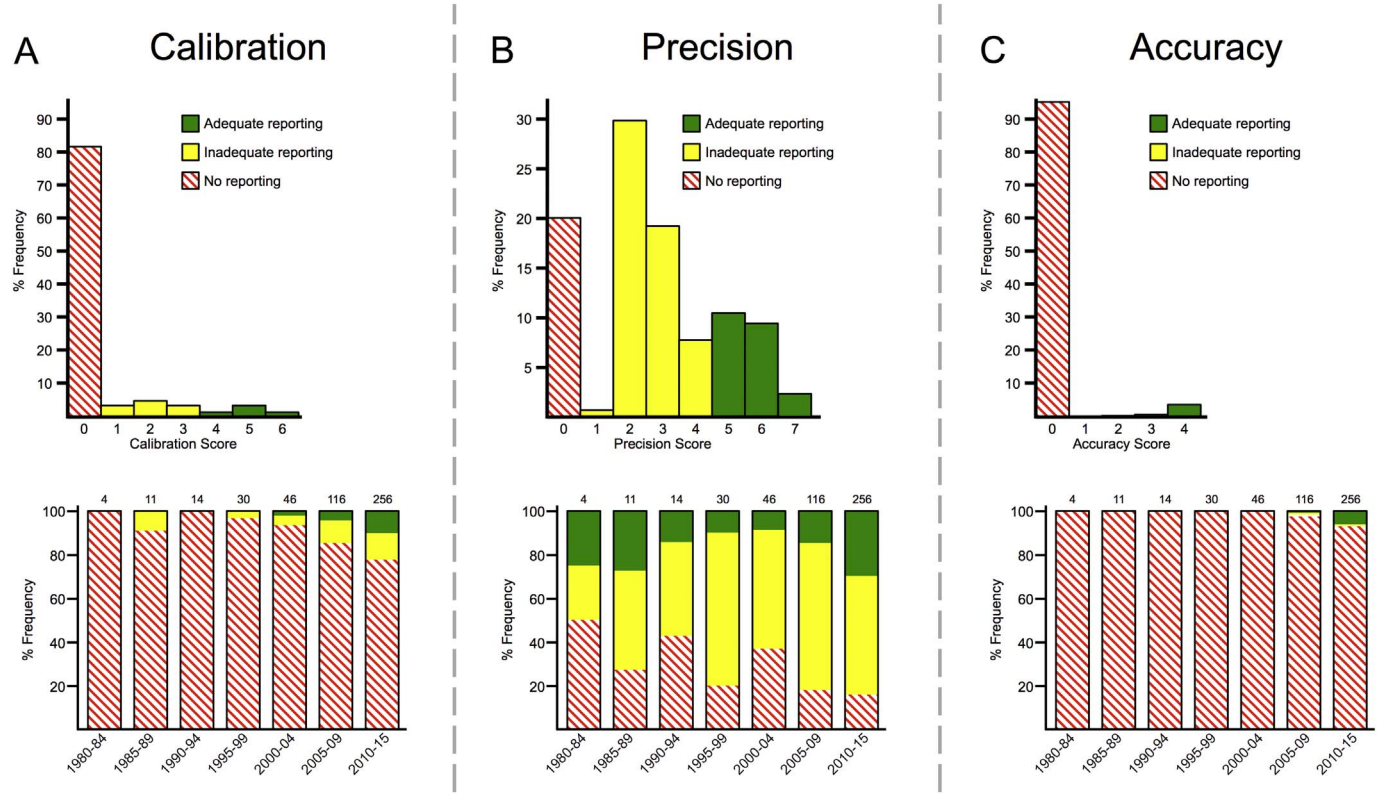

Fig. 3. Scores for (A) calibration, (B) precision, and (C) accuracy. The lower portion of the figures show the relative proportion of studies providing no information, insufficient information, and sufficient information for calibration, precision, and accuracy for each five-year interval.

several internal or international standards. In 20% of the papers surveyed, precision was linked to check standards that were directly associated with the samples analyzed in the paper, but the manner in which precision was quantified and linked to the check standards was clearly stated in only 11% of the papers surveyed. The check standards analyzed had similar elemental (wt% C and wt% N) compositions to the samples analyzed in 18% of the studies and dissimilar elemental compositions in 1% of the studies; the similarity in check standard elemental composition could not be evaluated in 81% of the studies. In other words, when the check standards were explicitly named they generally had similar isotopic compositions to the samples being analyzed. In the papers surveyed, precision was linked to check standards (24% of studies), sample replicates (7%), both check standards and sample replicates (4%), or calibration standards (1%). In 64% of studies, the method of quantifying precision was not specified even though a value was reported for analytical precision or error.

3.2.6. Sample replication

The analysis of replicate samples gives a marker of precision that is specific to the materials being analyzed. In most applications relevant to archaeology, the materials being analyzed (e.g., bulk collagen, hair, bulk plant material) are of a more heterogeneous nature than the materials used as standards, and sample processing techniques are employed as a means of homogenizing such samples (e.g., gelatinization of collagen, powdering of hair and plant samples). Standards, on the other hand, are often pure, single-chemical compounds (e.g., glutamic acid, alanine, sucrose) and are extremely homogeneous by nature. It was relatively common for studies to employ some level of sample replication (38% of studies explicitly reported replication), and in fact many studies (24%) analyzed 100% or nearly 100% of their samples in duplicate or triplicate. Despite this, it was rarely obvious what the level of sample-specific precision was, because no values relevant to sample replicates were actually reported (94% of studies failed to report any specific details) and it was simply stated that all of the samples were analyzed in duplicate or triplicate. When values were reported, 14% of studies (n = 6) reported the actual values of the sample replicates, 21% of studies (n = 9) reported the standard deviation of each sample replicate pair or set, 12% of studies (n = 5) studies reported the mean standard deviation of all of the replicates, and 53% of studies (n = 23) reported the average difference between duplicate pairs (note: some studies reported more than one of the above).

4. Discussion

As more studies are published it will become increasingly practical to conduct large-scale meta-analyses of particular regions or time periods involving thousands or perhaps tens of thousands of isotopic measurements. Moreover, complete datasets are being presented more regularly (Fig. 2A) and it is likely that the widespread opportunity to publish online-only supplemental files, and the rise of various online databases (Pauli et al., 2017), will facilitate even greater data disclosure. Comparability of data generated in different laboratories under different conditions therefore becomes a serious concern. Given the poor track record of reporting adequate quality control and quality assurance data in archaeology, it is critical that future studies make a concerted effort to make this a priority.

4.1. Why calibration matters

The methods by which isotopic measurements are calibrated relative to the international reference scales are exceptionally important. To demonstrate this, we present a series of carbon and nitrogen isotope measurements of standards with known isotopic compositions (Table 2) and perform a series of calibrations of these measurements using every possible single-point (calibration with one standard) and two-point (calibration with two standards) calibration. In Table 3 we present the equations used to calibrate the data (real data from a single

Table 2
Standard reference materials used in the example to demonstrate the importance of calibration.

Standard name	Material	Type	δ^{13}C (‰, VPDB)	δ^{15}N (‰, AIR)
USGS40	Glutamic acid	International	− 26.39 ± 0.04	− 4.5 ± 0.1
USGS41	Glutamic acid	International	+ 37.63 ± 0.05	+ 47.6 ± 0.2
MET	Methionine	Internal	− 28.60 ± 0.10	− 5.04 ± 0.15
SRM − 1	Caribou bone collagen	Internal	− 19.31 ± 0.12	+ 1.81 ± 0.12
SRM − 2	Walrus bone collagen	Internal	− 14.71 ± 0.13	+ 15.59 ± 0.10

analytical session) relative to the international reference scales (VPDB and AIR) for both two-point and single-point calibrations. Any standards that were not used in any given calibration were treated as check standards to assess measurement accuracy. The absolute differences of the observed relative to the known δ-values for the standards that were not used to calibrate the measurements are also presented in Table 3, such that smaller values indicate more accurate measurements and larger values indicate less accurate measurements. These data clearly demonstrate that two-point calibration curves produce more accurate results than single-point calibrations, with the check standards having δ^{13}C and δ^{15}N values that are consistently more different than their known values when calibrated with one standard relative to two standards (Table 3). Two-point calibration curves produce the most accurate results when the calibration standards have very different δ-values. For instance, the most accurate results are achieved when the data are calibrated using: USGS40 and USGS41 (δ^{13}C values differ by 64.0 ‰ and δ^{15}N values differ by 52.1 ‰), MET and USGS41 (δ^{13}C values differ by 66.2 ‰ and δ^{15}N values differ by 52.6 ‰), and SRM − 1 and USGS41 (δ^{13}C values differ by 56.9 ‰ and δ^{15}N values differ by 45.8 ‰). On the other hand, the results are far less accurate when the data are calibrated using USGS40 and MET (δ^{13}C values differ by 2.2 ‰ and δ^{15}N values differ by 0.5 ‰). Paul et al. (2007) demonstrated that there is no increase in accuracy with a multi-point (rather than two-point) calibration curve. While many of the 'error' values reported in Table 3 may seem small relative to some topics of archaeological interest (e.g., differences between C_3 and C_4 plants, trophic shifts), they could be propagated in comparisons of values obtained in different laboratories (or even in different analytical sessions), leading to much larger actual errors (Paul et al., 2007). Given that some calibration methods are clearly better than others, it is important the method used be explicitly stated in the materials and methods section.

4.2. Best practices for isotopic research in archaeology

Below we outline some recommendations for calibration methods, analytical accuracy, and analytical precision, based on the results of this study in combination with IUPAC guidelines and previous research (Carter and Fry, 2013; Coleman and Meier-Augenstein, 2014; Coplen, 2011; Paul et al., 2007; Skrzypek, 2013; Skrzypek et al., 2010; Werner and Brand, 2001). Additional consideration should be given to the manner in which the isotopic measurements themselves are reported and discussed, as the terminology used in the archaeological literature is frequently incorrect. A detailed treatment of this terminology is beyond the scope of this paper but the reader is referred to two recent papers covering precisely this topic (Bond and Hobson, 2012; Coplen, 2011).

Ideally, as much data as possible should be presented in reference to analytical uncertainty. However, for the sake of readability and to conform to length restrictions imposed by journals it may be impractical to report everything in the main body of the article. An alternative is to report an abbreviated discussion of analytical uncertainty in the

Table 3
Absolute difference in observed δ-values of standard reference materials with known isotopic compositions, calibrated using different methods. Each row in the table represents a different calibration curve using one or two of these standards. The values represent means for the differences of each standard (5–9 aliquots) relative to its known value. The values in the 'Mean' column represent the mean differences between all standards used as checks and their known values for each calibration. Values that are > 0.2 ‰ are in boldface and values > 0.5 ‰ are shaded. Note that the accuracy (calculated as the absolute difference between observed and known values) is far lower for single point calibrations and for two point calibrations where the δ-values of the two standards are similar (e.g., USGS40 and MET, see Table 2). No δ-value is reported for a standard when it was used for calibration and thus its observed value is by definition identical to its known value.

		Equations for Calibration Curves				Absolute Difference of Check Standards from Known Values (‰)													
		^{13}C		^{15}N		Mean		MET		SRM−1		SRM−2		USGS40		USGS41	
Calibration Standards		Slope	Intercept	Slope	Intercept	^{13}C	^{15}N	^{13}C	^{15}N	^{13}C	^{15}N	^{13}C	^{15}N	^{13}C	^{15}N	^{13}C	^{15}N
Two-point calibrations																	
USGS40	USGS41	1.016	−10.975	1.010	−1.155	0.05	0.08	0.03	0.01	0.04	0.12	0.08	0.10	–	–	–	–
USGS40	MET	1.004	−11.165	1.029	−1.092	**0.38**	**0.42**	–	–	0.12	<0.01	**0.22**	**0.28**	–	–	**0.79**	**0.99**
USGS40	SRM−1	1.022	−10.892	1.029	−1.093	0.13	0.41	0.04	<0.01	–	–	0.01	**0.27**	–	–	**0.35**	**0.97**
USGS40	SRM−2	1.023	−10.876	1.015	−1.138	0.15	0.12	0.04	0.01	0.01	0.08	–	–	–	–	**0.41**	**0.27**
MET	USGS41	1.016	−10.955	1.010	−1.164	0.03	0.08	–	–	0.01	0.12	0.05	0.11	0.03	0.01	–	–
MET	SRM−1	1.017	−10.928	1.029	−1.093	0.05	**0.62**	–	–	–	–	0.03	**0.27**	0.03	<0.01	0.10	**0.97**
MET	SRM−2	1.020	−10.887	1.016	−1.144	0.10	0.13	–	–	0.02	0.09	–	–	0.04	0.01	**0.25**	**0.28**
SRM−1	SRM−2	1.024	−10.869	1.009	−1.037	0.19	0.11	0.06	0.13	–	–	–	–	0.02	0.12	**0.50**	0.08
SRM−1	USGS41	1.015	−10.943	1.007	−1.032	0.03	0.10	0.02	0.14	–	–	0.04	0.02	0.04	0.13	–	–
SRM−2	USGS41	1.015	−10.906	1.007	−0.998	0.07	0.13	0.07	0.18	0.04	0.03	–	–	0.09	0.17	–	–
Single-point calibrations																	
USGS40	–	–	−11.219	–	−1.189	**0.36**	**0.25**	0.01	0.01	0.15	0.18	**0.26**	**0.30**	–	–	**1.01**	**0.52**
USGS41	–	–	−11.226	–	−1.203	**0.91**	**0.40**	**1.02**	**0.53**	**0.86**	**0.34**	**0.75**	**0.21**	**1.01**	**0.52**	–	–
MET	–	–	−10.208	–	−0.672	**0.37**	**0.26**	–	–	0.16	0.19	**0.27**	**0.32**	0.01	0.01	**1.02**	**0.53**
SRM−1	–	–	−11.069	–	−1.011	**0.32**	**0.21**	0.16	0.19	–	–	0.11	0.12	0.15	0.18	**0.86**	**0.34**
SRM−2	–	–	−10.960	–	−0.887	**0.35**	**0.24**	**0.27**	**0.32**	0.11	0.12	–	–	**0.26**	**0.30**	**0.75**	**0.21**

Materials and methods section of the text (see Section 5 for an example), with an expanded presentation of the relevant data in an Online Supplement (see Appendix E for an example).

For archaeologists not directly involved in instrumental analyses, the recommendations below should serve as a guide to what should be expected by a laboratory providing isotopic measurements. We suggest requesting full documentation of all analytical sessions (or 'runs'), including measured and calibrated values for all standards included in each run, and documentation of the calibration methods (see below). Calibration standards can be used to report precision (e.g., ± 1σ of measured values), but cannot give an indication of accuracy. Check standards can be used to report accuracy (based on the difference between observed mean values and known values) and precision (based on the standard deviation of repeated measurements). Replicate analyses of sample duplicates are the best measure of precision (based on mean absolute difference or mean standard deviations of replicate analyses), but cannot be used to evaluate accuracy.

4.3. Calibration recommendations

- A calibration curve should be generated for each analytical session using two SRMs (two-point calibration)
- The SRMs used to calibrate the data should be named and differentiated from SRMs that were not used to calibrate the data
- The δ-values of calibration SRMs should bracket those of the samples being measured
- The δ-values of the calibration SRMs that were used should be clearly stated, even if these were international SRMs with values that can be checked online. These values change periodically (e.g., Coplen et al., 2006; Qi et al., 2003) and it is therefore important to know precisely which values were used to calibrate the results
- The number of calibration standards used in each analytical session should be stated
- Each calibration standard must be analyzed multiple times in each analytical session and ideally 10% of the total analyses should be calibration standards

4.4. Monitoring systematic errors (accuracy)

- Accuracy cannot be determined using the δ-values obtained from calibration standards or sample replicates
- Check SRMs with known isotopic compositions should be analyzed during the same sessions as the samples being reported, but not included in the calibration curve
- Check SRMs should have δ-values within the range of values of the samples being analyzed and elemental compositions similar to that of the samples being analyzed
- The known δ-values (mean ± 1σ) of the check SRMs should be reported (see Section 5.0 and Appendix E)
  The known δ-values (mean ± 1σ) of the check SRMs should be reported (see Section 5.0 and Appendix E)
- The observed δ-values (mean ± 1σ, number analyzed) of the check SRMs should be reported. The mean δ-values for the check standards should be presented. Ideally, the means for each analytical session should be reported in the SOM (see Appendix E)
  The observed δ-values (mean ± 1σ, number analyzed) of the check SRMs should be reported. The mean δ-values for the check standards should be presented. Ideally, the means for each analytical session should be reported in the SOM (see Appendix E)
- A minimum of 10% of the total analyses in an analytical session should be check SRMs

4.5. Monitoring random errors (precision)

- Estimates of precision can be derived from the variability in results for check standards, calibration standards, and sample replicates (see Appendix E)
  Estimates of precision can be derived from the variability in results for check standards, calibration standards, and sample replicates (see Appendix E)
- The standard deviation of the δ-values should be reported for the check standards and calibration standards. Ideally, the means for each analytical session should be reported in an appendix (see Appendix E)
  The standard deviation of the δ-values should be reported for the check standards and calibration standards. Ideally, the means for each analytical session should be reported in an appendix (see Appendix E)
- For samples analyzed in duplicate, the pooled standard deviation of the δ-values for all duplicate pairs and the number of samples duplicated should be specified. Ideally, the actual δ-values of all the duplicated samples should be reported in an appendix (see Appendix E)
  For samples analyzed in duplicate, the pooled standard deviation of the δ-values for all duplicate pairs and the number of samples duplicated should be specified. Ideally, the actual δ-values of all the duplicated samples should be reported in an appendix (see Appendix E)
- For samples analyzed in triplicate (or greater), the pooled standard deviation of the δ-values of the sample replicate sets should be reported and the number of samples and rate of sample replication should be specified. Ideally, the actual δ-values of all the replicated samples should be reported in an appendix
- A typical analytical session should include a minimum of 10% sample duplicates but the rate of duplication will vary based on the heterogeneity of the samples being analyzed. Samples that are more heterogeneous (bulk plant tissues for example) should be analyzed at a higher rate of duplication than highly homogenous samples, such as bone collagen that has been solubilized

4.6. Reporting uncertainty

Although some percentage of studies surveyed reported some value related to uncertainty it was extremely rare that the method by which this value was obtained was specified. It is critical that the method used to calculate the reported uncertainty is clearly stated and a reference provided as appropriate. Following the guidelines presented by the Guide to the Expression of Measurement Uncertainty (Joint Committee for Guides in Metrology, 2008) and Nordtest (Magnusson et al., 2012) we present one such method for quantifying standard analytical uncertainty (Appendix F) that is straightforward and can be calculated using quality control and quality assurance practices (e.g., regularly analyzing check standards, analyzing samples in duplicate, triplicate, or x-licate) that should be performed by any reputable laboratory. A Microsoft Excel spreadsheet is provided that will allow for quick and easy calculation of analytical uncertainty (Appendix G). The sample IRMS dataset provided in Appendix A is used to illustrate how this spreadsheet is used.

5. Example paragraph on calibration and analytical uncertainty

The following paragraph provides an example of relevant details that should be included in studies presenting IRMS data. Although the example is for stable carbon and nitrogen isotopic compositions of bone collagen, the same general format can be applied to other comparable IRMS data. The data with which this paragraph is associated are presented in Appendix A.

Stable carbon and nitrogen isotopic compositions were determined using a Thermo Delta V continuous flow isotope ratio mass spectrometer coupled to a Costech ECS 4010 elemental analyzer at University X. Stable carbon and nitrogen isotopic compositions were calibrated relative to the VPDB and AIR scales using USGS40 and USGS41.

Measurement uncertainty was monitored using three in-house collagen standards with well-characterized isotopic compositions: IRM-1 (deer bone collagen, $\delta^{13}C - 19.28 \pm 0.07$ ‰, $\delta^{15}N + 1.79 \pm 0.11$ ‰), IRM-2 (sea lion bone collagen, $\delta^{13}C - 11.59 \pm 0.08$ ‰, $\delta^{15}N + 18.02 \pm 0.06$ ‰), and IRM-3 (cow bone collagen, $\delta^{13}C - 15.30 \pm 0.07$ ‰, $\delta^{15}N + 9.62 \pm 0.10$ ‰). Precision ($u(R_w)$) was determined to be ± 0.14 ‰ for both $\delta^{13}C$ and $\delta^{15}N$ on the basis of repeated measurements of calibration standards, check standards, and sample replicates. Accuracy or systematic error ($u(bias)$) was determined to be ± 0.07 for $\delta^{13}C$ and ± 0.11 for $\delta^{15}N$ on the basis of the difference between the observed and known δ values of the check standards and the long-term standard deviations of these check standards. Using the equations presented in Appendix F, the total analytical uncertainty was estimated to be ± 0.16 ‰ for $\delta^{13}C$ and ± 0.18 for $\delta^{15}N$. Additional details are provided in Appendix E.

6. Conclusions

An extensive review of papers presenting original isotopic results in archaeology revealed that isotopic data were typically presented with insufficient information about how δ-values were measured and calibrated, and how analytical uncertainty was calculated. To promote best practices among all who use and disseminate isotopic research, we have provided a series of recommendations for data analysis and reporting. Although we focused on carbon and nitrogen isotope compositions of organic materials, our general recommendations regarding monitoring and reporting analytical uncertainty are also applicable to other isotopic measurements, such as $\delta^{13}C$ and $\delta^{18}O$ in tooth enamel bioapatite. That said, each IRMS application presents its own unique set of challenges and may require even greater vigilance. For example, the measurement of $\delta^{2}H$ and $\delta^{18}O$ in archaeological hair will require a great deal more care with respect to monitoring memory effects and quantifying exchangeable H (Meier-Augenstein et al., 2013).

It is particularly important that researchers make calibration methods and uncertainty calculations explicit, utilize two-point (rather than single-point) calibrations, and monitor and report analytical accuracy using check standards that are not included in the calibration curve. Only if these guidelines are met can the full potential of isotopic research in archaeology be realized, through rigorous broad-scale comparisons of isotopic results obtained in different laboratories and from a diversity of temporal and geographical contexts.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jasrep.2017.05.007.

Acknowledgements

This research was supported by a Killam Postdoctoral Research Fellowship (JZM), an NSERC Banting Postdoctoral Fellowship (PS), a SSHRC Banting Postdoctoral Fellowship (JZM), and the Canada Research Chairs Program (PS).

References

Ambrose, S.H., 1990. Preparation and characterization of bone and tooth collagen for isotopic analysis. J. Archaeol. Sci. 17, 431–451.
Bond, A.L., Hobson, K.A., 2012. Reporting stable-isotope ratios in ecology: recommended terminology, guidelines and best practices. Waterbirds 35, 324–331.
Carter, J.F., Fry, B., 2013. Ensuring the reliability of stable isotope ratio data—beyond the principle of identical treatment. Anal. Bioanal. Chem. 405, 2799–2814.
Coleman, M., Meier-Augenstein, W., 2014. Ignoring IUPAC guidelines for measurement and reporting of stable isotope abundance values affects us all. Rapid Commun. Mass Spectrom. 28, 1953–1955.
Coplen, T.B., 1994. Reporting of stable hydrogen, carbon and oxygen isotopic abundances. Pure Appl. Chem. 66, 271–276.
Coplen, T.B., 2011. Guidelines and recommended terms for expression of stable-isotope-ratio and gas-ratio measurement results. Rapid Commun. Mass Spectrom. 25, 2538–2560.
Coplen, T.B., Brand, W.A., Gehre, M., Gröning, M., Meijer, H.A.J., Toman, B., Verkouteren, R.M., 2006. New guidelines for $\delta^{13}C$ measurements. Anal. Chem. 78, 2439–2441.
DeNiro, M.J., 1985. Postmortem preservation and alteration of *in vivo* bone collagen isotope ratios in relation to palaeodietary reconstruction. Nature 317, 806–809.
Guiry, E.J., Szpak, P., Richards, M.P., 2016. Effects of lipid extraction and ultrafiltration on stable carbon and nitrogen isotopic compositions of fish bone collagen. Rapid Commun. Mass Spectrom. 30, 1591–1600.
Jardine, T., Cunjak, R., 2005. Analytical error in stable isotope ecology. Oecologia 144, 528–533.
Joint Committee for Guides in Metrology, 2008. Evaluation of Measurement Data — Guide to the Expression of Uncertainty in Measurement. JCGM 100:2008. International Organization for Standardization, Geneva.
Jørkov, M.L.S., Heinemeier, J., Lynnerup, N., 2007. Evaluating bone collagen extraction methods for stable isotope analysis in dietary studies. J. Archaeol. Sci. 34, 1824–1829.
van Klinken, G.J., 1999. Bone collagen quality indicators for palaeodietary and radiocarbon measurements. J. Archaeol. Sci. 26, 687–695.
Magnusson, B., Näykki, T., Hovind, H.V., Krysell, M., 2012. Handbook for Calculation of Measurement Uncertainty in Environmental Laboratories. Nordtest Technical Report. 537 (ed. 3.1).
Mariotti, A., 1983. Atmospheric nitrogen is a reliable standard for natural ^{15}N abundance measurements. Nature 303, 685–687.
Meier-Augenstein, W., Hobson, K.A., Wassenaar, L.I., 2013. Critique: measuring hydrogen stable isotope abundance of proteins to infer origins of wildlife, food and people. Bioanalysis 5, 751–767.
Merritt, D.A., Hayes, J.M., 1994. Factors controlling precision and accuracy in isotope-ratio-monitoring mass spectrometry. Anal. Chem. 66, 2336–2347.
Ohlsson, K.E.A., Wallmark, P.H., 1999. Novel calibration with correction for drift and non-linear response for continuous flow isotope ratio mass spectrometry applied to the determination of $\delta^{15}N$, total nitrogen, $\delta^{13}C$ and total carbon in biological material. Analyst 124, 571–577.
Paul, D., Skrzypek, G., Fórizs, I., 2007. Normalization of measured stable isotopic compositions to isotope reference scales – a review. Rapid Commun. Mass Spectrom. 21, 3006–3014.
Pauli, J.N., Newsome, S.D., Cook, J.A., Harrod, C., Steffan, S.A., Baker, C.J.O., Ben-David, M., Bloom, D., Bowen, G.J., Cerling, T.E., Cicero, C., Cook, C., Dohm, M., Dharampal, P.S., Graves, G., Gropp, R., Hobson, K.A., Jordan, C., MacFadden, B., Pilaar Birch, S., Poelen, J., Ratnasingham, S., Russell, L., Stricker, C.A., Uhen, M.D., Yarnes, C.T., Hayden, B., 2017. Opinion: why we need a centralized repository for isotopic data. Proc. Natl. Acad. Sci. U. S. A. 114, 2997–3001.
Pestle, W.J., Crowley, B.E., Weirauch, M.T., 2014. Quantifying inter-laboratory variability in stable isotope analysis of ancient skeletal remains. PLoS One 9, e102844.
Prosser, S.J., 1993. A novel magnetic sector mass spectrometer for isotope ratio determination of light gases. Int. J. Mass Spectrom. Ion Process. 125, 241–266.
Qi, H., Coplen, T.B., Geilmann, H., Brand, W.A., Böhlke, J.K., 2003. Two new organic reference materials for $\delta^{13}C$ and $\delta^{15}N$ measurements and a new value for the $\delta^{13}C$ of NBS 22 oil. Rapid Commun. Mass Spectrom. 17, 2483–2487.
Sealy, J., Johnson, M., Richards, M., Nehlich, O., 2014. Comparison of two methods of extracting bone collagen for stable carbon and nitrogen isotope analysis: comparing whole bone demineralization with gelatinization and ultrafiltration. J. Archaeol. Sci. 47, 64–69.
Skrzypek, G., 2013. Normalization procedures and reference material selection in stable HCNOS isotope analyses: an overview. Anal. Bioanal. Chem. 405, 2815–2823.
Skrzypek, G., Sadler, R., Paul, D., 2010. Error propagation in normalization of stable isotope data: a Monte Carlo analysis. Rapid Commun. Mass Spectrom. 24, 2697–2705.
Werner, R.A., Brand, W.A., 2001. Referencing strategies and techniques in stable isotope ratio analysis. Rapid Commun. Mass Spectrom. 15, 501–519.