Quaternary International

Best practices for selecting samples, analyzing data, and publishing results in isotope archaeology

--Manuscript Draft--

Manuscript Number:	QUATINT-D-21-00496R2
Article Type:	SI: Multiple Scales of Stable Isotope Palaeoecology_closed
Keywords:	Isotope archaeology; Palaeodiet; palaeoclimate; Ancient mobility; Quality control; P-values
Corresponding Author:	Petra Vaiglova Washington University In St Louis: Washington University in St Louis St Louis, UNITED STATES
First Author:	Petra Vaiglova, PhD
Order of Authors:	Petra Vaiglova, PhD
	Nicole A. Lazar, PhD
	Elizabeth A. Stroud, PhD
	Emma Loftus, PhD
	Cheryl A. Makarewicz, PhD
Abstract:	Isotopic analysis has become one of the most popular arenas of archaeological science, in part due to its versatility to uncover intriguing insights from a range of organic and inorganic archaeological materials. However, alongside an increase in popularity, the field has seen the rise of dissemination of publications that do not pass quality control, do not apply robust interpretative frameworks, or do not report data in ways that would make them amenable to critical evaluation or inclusion in large meta-analyses. This paper represents an effort to clarify some of the most pressing weaknesses and misconceptions in 'traditional' applications of isotopic techniques in archaeology: measurement of stable carbon, nitrogen, and oxygen isotope values of organic and inorganic materials (bulk bone collagen, bulk dentine, and seeds; bulk and incremental tooth enamel, and molluscan shells), and strontium isotope ratio analysis of tooth enamel and cremated bone. The discussion centers on three key aspects of research: (1) Selecting samples, with advice on building comparative baselines (or more appropriately 'base intervals ') and words of caution on interpreting stable carbon isotope values measured during AMS radiocarbon dating. (2) Handling data, including tips on exploratory data analysis, graphical visualization, and statistical assessment of differences between groups; with particular reference to the Statement on p -values published by the American Statistical Association. (3) Reporting data, with advice on using correct terminology and decimal points, calculating measurement precision and accuracy, and communicating results using effective scientific language. The advice provided in this paper does not cover all aspects of project design and dissemination, but will hopefully provide clarification within these three key areas and inspire further discussion of effective and impactful applications of isotopic techniques in archaeology.
Suggested Reviewers:	Tamsin O'Connell, PhD Professor of Archaeological Science, University of Cambridge Central Science Library: University of Cambridge tco21@cam.ac.uk Through her work, Prof O'Connell has advocated for greater scientific rigor and publications standards in isotopic archaeology.
	Paul Szpak, PhD Assistant Professor in Environmental Archaeology, Trent University paulszpak@trentu.ca Prof Paul Szpak has published on applications of rigorous quality control criteria and robust assessment of measurement error.

Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation

	Eric Guiry, PhD Lecturer in Biomolecular Archaeology, University of Leicester ejg26@leicester.ac.uk Dr Guiry has published numerous works on quality control and robust applications of isotopic methods in archaeology and paleoecology.
	Diana Moreiras Reynaga, PhD Honorary Research Associate, The University of British Columbia dianakarina.moreiras@ubc.ca Dr Moreiras Reynaga has published excellent work in the area of building robust ecological isoscapes.
	Kate Britton, PhD Reader in archaeological geoscience, University of Aberdeen k.britton@abdn.ac.uk Dr Britton has published an influential and very thorough review of isotopic analyses of tooth bioapatite and has a long publication list of quality applications with archaeological geosciences.
	Sarah Pederzani, PhD Postdoctoral Researcher, Max-Planck-Institute for Evolutionary Anthropology: Max-Planck-Institut fur evolutionare Anthropologie sarah_pederzani@eva.mpg.de Dr Pederzani published a very thorough review of isotopic analysis of tooth bioapatite and has contributed major papers on applications of isotope analyses to ancient paleoclimate reconstructions.
	Anneke Janzen, PhD Assistant Professor in Anthropology, UT Knoxville: The University of Tennessee Knoxville ajanzen@utk.edu Dr Janzen has published very robust applications of isotope techniques in archaeological paleomobility and animal management studies.
	Jessica Metcalfe, PhD Assistant Professor in Archaeological Science, Lakehead University jmetcal1@lakeheadu.ca Dr Metcalfe has contributed to an important paper on reporting measurement error in isotope archaeology and has a demonstrated record of quality publications in the field.
	Matthew Collins, PhD Professor in Paleoproteomics, University of Cambridge matthew@palaeome.org Prof Collins has a very long track record of outstanding research in archaeological science. He has collaborated on numerous isotopic studies and can provide a wide-scientific perspective on the issues raised in the submitted work.
Response to Reviewers:	

Prof Jule Xiao
Editor-in-Chief
Quaternary International

January 31, 2022

Dear Prof Xiao,

Please find enclosed our revised manuscript titled "Best practices for selecting samples, analyzing data, and publishing results in isotope archaeology", which we wish to submit for publication in the Special Issue *Multiple Scales of Stable Isotope Palaeoecology*.

We have carefully gone through the suggestions from the reviewers and respond to all their queries. Many of the 'tracked changes' represent stylistic updates that we made to improve the readability and the flow of the text. The only content-specific changes that were made relate to the issues raised by the reviewers and referenced with line numbers in the attached "Response to Reviewers file."

We also wish to inform you that two previously listed authors, Dr Rick Schulting and Dr Melanie Beasley, have asked for their names to be removed because they felt they have not made a substantial enough contribution to the manuscript. Because the comments that they provided during the writing process, we included their names in the Acknowledgements.

Thank you very much for consideration and please don't hesitate to get in touch with me with any questions.

On behalf of all the authors,

Petra Vaiglova

Postdoctoral Research Associate
Department of Anthropology
Laboratory Coordinator
Laboratory for the Analysis of Early Food-Webs (LAEF)
Washington University in St. Louis
Phone: 860-634-0676
Email: petra.vaiglova@wustl.edu

Manuscript Number: QUATINT-D-21-00496R1

Best practices for selecting samples, analyzing data, and publishing results in isotope archaeology

Reviewer #1:
The authors have made major revisions to the original article in response to the reviewer comments. The discussion on baseline variability has much improved. I now find the aspects covered to be more nuanced and clear and with more actionable advice compared to the previous version. I'm sad to see that the section on designing research questions has been removed completely, as I think the topic of study and hypothesis design is quite important, but I understand that time-constraints are preventing the authors from fully reworking the section. I also find that instances of incorrect isotopic terminology or concepts have been corrected and citations have been added throughout the article. I also appreciate the effort that the authors have made to make the tone of the paper more neutral and have not found more instances of language that I could see to be read as judgmental. My comment on multiple testing was I believe misunderstood by the authors, but the reworking of the paper has removed my issues anyway, so it is a moot point to argue this further.
I found some minor things in the revised version (noted below), but these should be easily fixable.

Line 439-440: Due to the extended nature of enamel mineralization, the amount of time averaged in a single sample is very poorly constrained for almost all species commonly used for incremental sampling of tooth enamel, but likely exceeds this time estimate. Please revise with more recent citations.
>> The whole paragraph was revised (lines 876–891 in version with tracked changes)

Line 444 ff.: I suggest the authors clearly separate considerations pertaining to bulk samples and incremental samples. Sampling a fixed distance is most important for bulk sampling, where seasonal bias cannot be identified from the obtained measurements. Although it should be noted that some species (e.g. rhinos) don't have fixed tooth development times, seasonally speaking, and noise from seasonal bias will be random and could be compensated for using a larger sample size. These concerns don't apply to sequential data in the same way in my opinion. While obtaining a full annual cycle is desirable, its absence is clearly visible in the data obtained. At the same time, whether such teeth have to be excluded from analyses depends on the type of analysis. Sequential data can for instance be analysed by extracting and comparing summer and winter values, which allows for the inclusion of data points from individuals with less than a full year of input recorded. I suggest revising this paragraph accordingly.
>> We disagree with the reviewer on this point. Because partial sequences are truncated, it is often impossible to determine which seasons are actually represented. In addition, because they do not necessarily capture the real maximum, minimum or average intra-tooth values (since many values are missing), it would be inappropriate to use them in inter-teeth comparisons. We clarified this in the text (lines 894–901) and included Supplementary Material 2 to further discuss and illustrate the point.

Line 485: A word is missing in this sentence.
>> Fixed

Line 487: I would rephrase this, as the order of the sentence could be misunderstood to mean that $\delta^{13}C$ adds mobility information. Maybe something like 'useful for obtaining additional information on dietary ecology' would be clearer. I would also caution against suggesting that $\delta^{18}O$ is primarily useful for intensifying mobility given the low spatial resolution this tracer offers by itself.
>> Fixed (now on line 1010–1012)

Line 650 ff. A word seems to be missing from this sentence. Given the palimpsest nature of many archaeological deposits it may be advisable to phrase this as 'as contemporary as possible'.
>> Fixed (now on line 1291)

Line 1054: I'm not sure why the authors chose this phrasing. On the one hand, the inclusion of these conventions in this paper seems to suggest that the authors recommend following them, on the other hand they describe them as 'often dismissed as too pedantic'. If the authors think these conventions are not worth following, why are they spending a whole paragraph discussing them? I would suggest rephrasing to make clear what the goal of this paragraph is.
>> Fixed (now line 2284–2290)

Reviewer #2:
This paper provides useful guidelines for archaeologists using stable isotope analysis in their research. The authors cover several areas in where archaeologists commonly make mistakes, or where writing and presentation of data could be improved. This includes assessing sample integrity, exploratory data analysis, formal analysis, and finally some pointers on presenting/discussing results and publication. The authors have made substantial changes since the first round of review, which has improved the manuscript substantially. The manuscript is generally clear and well-organized. My only major concern is Section 5.4: I think that this is a worthy topic, but I'm not sure how effective it is given the space constraints in the paper, particularly without specific examples of what this kind of writing should look like. Finally, the manuscript would benefit from some light editing for formatting, there are a few extra spaces here and there.
>> We feel very strongly about including section 5.4, and included Supplementary Material 3 to provide examples from isotope archaeology. Graduate students in archaeology often do not get sufficient training in how to write, and we feel that including this section will be helpful for getting them (and other scientists) to think about some of the issues that come up in writing.

For Section 3, I found the addition of example graphs in the SI useful.
>> Thank you

Line 36 ff: I don't quite understand the wording "marine/molluscan shell": molluscs (which include bivalves like clams and gastropods like snails) can be found in marine and freshwater environments, and individuals from both environments can be analyzed for their stable isotope composition. Just write "molluscan shell"

>> fixed

line 166: change “less” to “fewer”
>> changed

line 169: explain more clearly what you mean by “worst case scenario”
>> Reworded (line 582–585)

line 202 paragraph: I like this addition. I would add here that it is important for archaeologists to also include tables of element data (element, side, and element portion, if necessary, as well as any other relevant information that allows them to show that a specimen is from a different individual than the other specimens analyzed) in their studies (and if they do not know how to identify these, they should consult their friendly neighborhood zooarchaeologist!). This should be reiterated in the conclusion as well. In this paragraph I would also add some references to further illustrate this point (e.g. Balasse et al. 2013 European Journal of Archaeology 16(2):221-246…I see you cite this one already!).
>> Thank you for the suggestion. We added a stentence on this, on line 699–702.

Line 206: “selecting the bones of a single element [insert: “for each taxon”], taking into account…”
>> added

Line 207: Here I would specify that the archaeologists selected the most abundant element from each taxon.
>> fixed (line 693)

Line 231: I would reword the text in the parentheses to say “heating at high temperatures, e.g. cremation”
>> Reworded as “heated to temperatures that turn them white”, because we are not talking about all heating temperatures, but only a specific range of temperatures that causes bones to become calcined (line 784)

Line 251: teeth of smaller taxa, like caprines, are usually sequentially sampled with negligible space between each sample.
>> We updated the section on sampling resolution (lines 876–891)

Line 253: crown height (not length)
>> fixed

Line 263: this sentence reads a bit unclearly.
>> fixed, now line 905–931

Sentence starting line 266: do the authors mean “minimum” instead of “pre-determined”?
>> No, we mean pre-determined (i.e. you decide ahead of time how long the sampling distance will be). We used the term “fixed distance” instead (line 931)

Line 278: replace “herbivore” with “bovid”
>> the reference to bovids was removed through rewording of the sentence (line 943–944)

Line 281: also cite here Fremondeau et al. 2012. doi:10.1016/j.jas.2012.04.0 04
>> Added

Line 297–298: word missing in this sentence?
>> Fixed (line 1009)

Line 300: add “seasonality” to the list of questions that can be explored here.
>> Added (line 1011–1012)

Line 318: Perhaps the authors could replace “inter-plant” variability with inter-crop? Could bulk samples include grains from multiple years?
>> fixed inter-crop. Yes, bulk samples could include grains from multiple years, which we specify on line 1017.

Line 326: the authors should start off this paragraph by noting that most seeds recovered from archaeological contexts are charred.
>> done and discussion on charring added (lines 1085–1092)

Paragraph starting line 326: What’s the range of temperatures that seeds can survive? Do we often find seeds that were heated to higher temps? Or do we find unburned seeds? Similar to how the authors outlined the kinds of conditions that favor preservation for skeletal materials, it would be good to include a note about this here.
>> Clarified in new paragraph (lines 1085–1082)

Line 340: citation for study that includes cross-section photos?
>> No one has done this yet, and we included a mention of both supplementary materials and open access databases.

Line 353: also include freshwater molluscs.
>> Changed to aquatic and terrestrial (line 1144)

Sentence starting line 364: should also include citations of studies that used data from modern molluscs
>> The paragraph was updated and Galimberti et al. (2017) reference was added (1156–1162)

379-380: does this also include freshwater molluscs?
>> Yes, changed (line 1193)

Line 393: end parenthesis is missing.
>> fixed

Line 447: “reversal of the canopy effect” reads a bit unclearly for me. Perhaps word a different way?
>> reworded (line 1304–1305)

Line 476: wetlands should be plural
>> fixed

Line 551: this example is a bit unclear to me. Perhaps reword or clarify.
>> reworded (line 2028–2031)

Example in 4.1: I think it would be useful to include the figures here.
>> both Fig.1 and 2 belong to section 4.1

Best practices for selecting samples, analyzing data, and publishing results in isotope archaeology ~~sample selection, data analysis, and publication of results in isotope archaeology~~

Petra Vaiglova[a,*], Nicole A. Lazar[b], Elizabeth A. Stroud[c], Emma Loftus[d], Cheryl A. Makarewicz[e] ~~Melanie Beasley[f], Rick J. Schulting[e]~~

[a] Department of Anthropology, Washington University in St Louis, 1 Brookings Drive, St. Louis, MO, 63130, USA
[b] Department of Statistics, Penn State University, 326 Thomas Building, University Park, PA, 16802, USA
[c] School of Archaeology, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, UK
[d] Institute for Pre- and Protohistoric Archaeology and Provincial Roman Archaeology, Ludwig Maximilian University Munich, Geschwister-Scholl-Platz 1, 80539 Munich, Germany
[e] Institute of Prehistoric and Protohistoric Archaeology, University of Kiel, Johanna-Mestorf-Straße 2-6, Kiel, Germany
~~[f] Department of Anthropology, Purdue University, West Lafayette, IN 47907, USA~~

* Corresponding author.
E-mail address: vaiglova.petra@gmail.com (P. Vaiglova).

Keywords

Isotope archaeology; Palaeodiet; Palaeoclimate; Ancient mobility; Quality control; P-values

Abstract

Isotopic analysis ~~of archaeological materials~~ has become one of the most popular arenas of archaeological science, ~~given,~~ in part~~,~~ due to its versatility to uncover intriguing insights from a range of organic and inorganic archaeological materials. However, alongside an increase in popularity, the field has seen the rise of dissemination of publications that do not ~~materials including human and animal remains, charred and desiccated plant grains and seeds, and climatic archives such as molluscan shells. Along with its rising popularity, however, the field has seen dissemination of publications that do not necessarily~~ pass quality control, do not apply robust interpretative frameworks, or do not report data in ways that would make them amenable to critical evaluation or inclusion in large meta-analyses ~~fully present data that can be critically evaluated or compared in larger meta-analyses~~. This paper represents an ~~collective~~ effort to clarify some of the most pressing weaknesses and misconceptions in 'traditional' applications of isotopic techniques

in archaeology: measurement of stable carbon, nitrogen, and oxygen isotope values of organic and inorganic materials (bulk bone collagen, bulk dentine, and seeds; bulk and incremental tooth enamel, and molluscan shells), ~~some of the weaknesses and misconceptions in applications of the 'traditional' approaches in isotopic archaeology: measurement of bulk stable carbon, nitrogen, and oxygen isotope values of organic (bone and dentine collagen, charred plant remains) and inorganic (tooth enamel, marine/molluscan shell) materials,~~ and strontium isotope ratio analysis of tooth enamel and cremated bone. The discussion centers on three key aspects of research:

(1) ~~s~~Selecting samples~~ample selection~~, with advice on building comparative baselines (or more appropriately 'base*intervals*') and words ~~of bones, teeth, plant grains/seeds, molluscan shells, with~~ of caution on interpreting stable carbon isotope values measured during AMS radiocarbon dating. ~~comments on interpreting stable carbon isotope values measured during AMS radiocarbon dating and advice on building comparative 'baselines' (or more appropriately 'baseintervals');~~ (2) ~~d~~Handling data~~ata handing~~, ~~-~~including tips on ~~-~~exploratory data analysis, graphical visualization, and statistical ~~data visualization, statistical~~ assessment of differences between groups~~,~~; ~~with~~ with particular reference to the Statement on *Pp*-values ~~released~~ published by the American Statistical Association~~; and~~.

(3) Reporting data~~data reporting~~, with advice on using correct terminology and decimal points, calculating measurement precision and accuracy, and communicating results using effective scientific language. ~~advice on the correct use of isotopic terminology, appropriate measurement calibration methods, complete assessment of measurement error, and effective use of language in scientific writing.~~

The advice provided in this paper does not ~~encompass~~ cover all aspects of project design and dissemination, but will hopefully provide clarification within these three key areas and inspire further discussion of effective and impactful applications of isotopic techniques in archaeology. ~~, but we hope will provide clarification where needed, and inspire further discussion about effective and impactful application and method development.~~

1. Introduction

Since J.J.Thomson et al. (1921) first discussed the concept of elemental isotopes 100 years ago, isotopic analyse~~i~~s ha~~s~~ve gained traction in a wide array of fields including geochemistry, climatology, hydrology, ecology, plant physiology, palaeobiology, bacteriology, and archaeology

Formatted: Font: Not Italic

Formatted: Font: Italic

(Sharp 2017). TPioneering he first stable isotope research in archaeologyisotopic analyses of archaeological remains in the 1970s aimed at identifying human consumption of non-native maize in a region where maize was not nativeNorth America (Vogel and van der Merwe 1977). Not long after, Ericson (1985) proposed that strontium isotope ratios of strontium could be used to trace the to infer the geographical origins of human remainss from archaeological assemblages. However, it was not until foundational strides were made towards had been made in understanding tissue diagenesis and preservation of *in vivo* isotopic composition of archaeological materials (e.g. Sullivan and Krueger 1981; Lee-Thorp and Van der Merwe 1987; Masters 1987; Schoeninger et al. 1989), paralleled byalongside advances in understanding biomolecular biochemical mechanisms such as routing of dietary isotopes (e.g., Ambrose and Norr 1993) and incorporation of strontium into mineral bioapatite (e.g., Rokita et al. 1993), that isotopee archaeology was on track to becoming one of the most popular branches of archaeological science. Today, its wide-ranging applications encompass both more 'traditional' and rather inexpensive analyses of bulk stable isotope values (e.g., bulk bone collagen, bulk and sequential incremental tooth enamel), as well as more instrumentally challenging and costlier high-resolution analyses (e.g., compound specific stable isotope analysis of bone collagen amino acids, high-resolution elemental mapping of tooth cross-sections on the microscopic level).

The growth of isotope archaeology was in large part driven by a generation of women scientists including Julia Lee-Thorp, Judith Sealy, Margaret Schoeninger, Noreen Tuross, Marilyn Vogel, and M. Anne Katzenberg. , who, tThrough their perseverance and adherence to high standards of scientific rigor, these scientists trained a long list of students and produced work and trained scientists that would come to transform the way we ask questions about the human past. Forty five years after the first application of stable isotope analysis in archaeologybirth of isotope archaeology (Vogel and van der Merwe 1977), the core principles that characterize these researchers' work—asking meaningful questions, identifying reliable samples, building robust interpretative frameworks, and producing scientifically rigorous publications—remain the cornerstones of impactful isotopic research.research in this field.

This paper forms part of a Special Issue dedicated to the career and mentorship of Julia Lee-Thorp and aims to synthesize useful advice on obtaining and working with isotopic data in archaeological

research. ~~This paper forms part of a Special Issue dedicated to Julia Lee-Thorp~~Through ~~,~~ ~~who~~collaboration with archaeologists and palaeoecologists, Julia Lee-Thorp made foundational strides towards understanding the outcome and the process of tissue diagenesis and applying isotopic analyses for understanding early ~~through collaboration with archaeologists and palaeoecologists in characterizing the process and outcome of tissue diagenesis, and~~hominin diets and past human–ecology dynamics. ~~in applying the rapidly evolving isotopic techniques to understanding aspects of early hominin diets and past human–ecology dynamics. To honor Julia Lee-Thorp's career to date, this paper aims to synthesize useful advice for isotope practitioners on obtaining and working with isotopic data in archaeological research.~~ As it is not possible ~~in one paper~~ to cover in ~~more than superficial~~sufficient detail all aspects of project design ~~for all the major themes addressed in isotopic archaeology~~ in one paper, we have chosen to focus on three key areas: ~~(a)~~ sample selection, ~~(b)~~ data processing ~~(exploratory data analysis, data visualization, statistical assessment of differences between groups)~~, and ~~(c)~~ data reporting ~~(terminology, measurement calibration, measurement error, decimal points, writing with clarity and brevity)~~. The ~~paper~~ discussion ~~is primarily~~ focuse~~d~~ on the ~~more~~ most common~~ly~~ ~~applied as well as misused~~ applications: ~~bulk~~ stable carbon, nitrogen, and oxygen isotope values of organic and inorganic materials (bulk bone collagen, bulk dentine, and seeds; bulk and incremental tooth enamel, and molluscan shells), and strontium isotope ratio analysis of tooth enamel and cremated bone. ~~stable carbon, nitrogen, and oxygen isotope analyses of organic (bone and dentine collagen, charred plant remains) and inorganic (tooth enamel, marine/molluscan shell) materials, and strontium isotope ratio analysis of tooth enamel and cremated bone.~~ The paper ~~It~~ does not aim to expose ~~blunders~~ errors by any particular researchers, nor does it try to generalize about trends in the amount of biochemical and geochemical training ~~trends in geochemical training~~ that practitioners of isotope archaeology receive. It is simply aimed at outlining where improvements can be made in the production of meaningful and interesting research.

In some parts, this paper reiterates advice presented in Roberts et al.'s (2017a) '*Calling all archaeologists: guidelines for terminology, methodology, data handling, and reporting when undertaking and reviewing stable isotope applications in archaeology'* and Szpak et al.'s (2017) '*Best practices for calibrating and reporting stable isotope measurements in archaeology*'. This

paper is meant to serve as a companion to these publications (and others including Fry 2006; Bond and Hobson 2012; Sharp 2017), while providing some more recent advice on:

- Assessing sample quality for collagen (Guiry and Szpak 2021), charred seeds (Szpak and Chiou 2019), bioapatite (France et al. 2020) and shell (Loftus et al. 2015)
- ~~Inspecting both surface and internal structure of grains to select charred plant material likely to provide reliable isotopic measurements .Applying more recently determined quality control criteria~~ (Loftus et al. 2015)
- ~~Sampling charred plant remains for isotopic analysis, in light of more recent research on using both surface and internal characteristics of grains to assess their suitability to isotopic analysis (Elizabeth~~ Stroud et al. in prep)~~.~~
- Identifying archaeological reference samples with high likelihood of approximating the unaltered local isotopic baseline (or base*interval*) values.
- Determining the suitability of stable carbon isotope values measured during AMS radiocarbon dating for palaeodietary and palaeoclimate research.~~Using stable isotope measurements produced during AMS radiocarbon dating~~
- Effectively using exploratory data analysis and graphical visualization techniques to guide statistical analysis and effective presentation of data.
- ~~Reporting statistical analyses in line with the Data visualization and exploratory data analysis~~
- ~~Carrying out and reporting statistical analyses in isotope archaeology, following the~~ American Statistical Association's recent Statement on ~~*pp*~~-values (Wasserstein and Lazar 2016; Wasserstein et al. 2019). ~~An example is provided showing how to assess and report meaningful differences between isotopic values of distinct groups.~~

For a review of ~~the method and~~ basic principles and research questions in isotope archaeology, we refer the reader ~~is referred~~ to Julia Lee-Thorp's paper ~~:~~*On isotopes and old bones*~~.~~ (2008).

Formatted: Font: Italic

2. Identifying reliable samples ~~that will provide reliable and relevant data~~

A mass spectrometer will always produce a measurement, no matter the sample preservation or the technical expertise of the operator. For this reason, it is important that~~ prior to analysis~~ researchers invest time in selecting samples that are likely to have retained ~~show a likelihood of~~

~~retaining~~ their *in vivo* (i.e. original) isotopic composition. ~~Here~~ In this section, we discuss considerations of sample size, sample type, sampling contexts, and preservation potential for the main materials analyzed in isotope archaeology: bones, teeth, ~~charred plant material~~seeds, and shells.

Equally important as selecting reliable samples is to confirm that the acquired measurements are reliable. This can be accomplished, with caveats, by assessing the degree to which the chemical composition of the archaeological samples resembles the composition of analogous modern material ~~comparing the chemical composition of the archaeological samples to those of modern samples that have been used to establish quality control (QC) benchmarks~~ (Zazzo et al. 2004; Beasley et al. 2014; Loftus et al. 2015; Szpak and Chiou 2019; Guiry and Szpak 2021). Table 1 provides a summary of currently accepted or suggested quality control (QC) benchmarks ~~relevant QC criteria and currently accepted/suggested benchmarks~~ for ~~bone~~ collagen, bone bioapatite, tooth enamel bioapatite, charred plant remains, and molluscan shell. It is important to recognize, however, that these criteria are not perfect, and will continue to be refined through further development (see discussion in Cheung and Szpak 2021).

It is also important to note that sample selection requires establishing a compromise between ~~acquiring~~ adequate sample size~~s to provide archaeologically meaningful datasets~~ and ethical responsibility to preserve ~~considerations to limit destructive sampling of~~ irreplaceable archaeological materials (Pálsdóttir et al. 2019; Squires et al. 2019). Curators are professionally obliged to safeguard archaeological collections~~,~~ and deny access to large sample numbers that may ~~where ever larger sample sizes may~~ offer little in the way of improved resolution. To justify destructive sampling, researchers should outline reasonable expectations surrounding the isotopic variability in the local ecology and support their desired numbers of samples with statistically meaningful estimates of optimal sample size (carried out using, for example, power analyses). In addition~~Well-defined statements of the archaeological question that integrate understanding of the isotopic variability of the system in question, alongside a grounding in statistical analytical methods (e.g. power analysis, see below), will help to justify appropriate sampling strategies that have a greater chance of achieving the project goals. Further~~, intentions to ~~motivations for archaeological~~ sampl~~e~~ing ~~of~~ human remains must not only comply with relevant laws and

Formatted: Font: Not Italic

regulations, but demonstrate the researchers' commitment to responsibly handle ~~explicitly address additional ethical considerations for~~sensitive material and involve local communities in the research, particularly when it relates to descendant groups or controversial political discussions (Atalay 2010; Roberts et al. 2018; Collard et al. 2019). ~~these sensitive materials (especially where investigations relate to modern groups or political discussions).~~

2.1 Bones

Sampling bones for palaeodietary studies needs to strike a ~~fine~~ balance between the desired level of detail and the minimum number of samples that ~~Minimum Numbers of well-preserved Individuals (MNI) that~~ will provide a 'best case scenario' for approximating the means and variability within a given population or group ~~approximating the population or the group mean~~. The ~~desired~~ level of attainable detail is ~~delineated~~ constrained by the research question and the ~~sources of~~ isotope variability characterizing ~~in the~~ the relevant ~~given foodweb~~ecosystem. For example, a project that aims to assess the overall dietary composition and homogeneity of a given population and a project that aims to assess possible differences in dietary intake between distinct sub-groups (identified by, for example, sex, age, pathologies, or association with distinct burial types, household units or site clusters) ~~if the aim is to understand the overall composition and homogeneity of diets of a given community, and compare this inference to the overall composition and homogeneity of diets at other sites, the sampling will need to focus on obtaining a representative coverage of people from the site's range of spatial and temporal contexts; with sample sizes appropriate for approximating the averages values and intra-site variability~~ (e.g. Müldner and Richards 2007) will delineate different optimal sample sizes. ~~If the aim is to assess differences in dietary inputs between meaningful groups such as males, females, children, individuals with certain pathologies, individuals buried in distinct types of burials, or individuals associated with separate household units or site clusters, the sampling will need to focus on obtaining large enough sample sizes from each context of interest that will enable an approximation of the averages and variability within each group (e.g. Pearson et al. 2015).~~

Power analysis can be used to estimate the minimum number of samples required to answer a research question (Murphy et al. 2014). It can be used to perform sample size calculations for a range of reasonable standard deviations in order to determine how much the outcomes change as

a result. This information can be used to establish the optimal sample numbers that will be amenable for answering the research question while avoiding redundancy (Pechenkina et al. 2005; Stevens et al. 2006). The indices required to carry out these calculations (likely effect sizes, group differences, and measures of variability such as standard errors) can be obtained from the literature or through pilot studies. They do not have to be precise, since the goal is to obtain a general idea of the sample sizes needed to detect, for example, differences of a specified magnitude.

In their paper *Counting sheep: sample size and statistical inference in stable isotope analysis and palaeodietary reconstruction*, Pearson and Grove (2013) carried out statistical simulations to determine the optimal sampling size for a large dataset of sheep (n=174) whose dietary intake was influenced by a combination of anthropogenic and environmental influences. The results showed that when using at least 8 samples per group, the 'true mean' of the group was represented in over 90 % of the cases. When using fewer than 8 samples per group, the 'true mean' was often not captured

~~In their paper *Counting sheep: sample size and statistical inference in stable isotope analysis and palaeodietary reconstruction*, Pearson and Grove (Pearson and Grove 2013). demonstrated using statistical simulations that when using less than 8 samples per subset, the 'true mean' of the population was often not captured, while when using at least 8 samples, it was represented in over 90 % of cases. This was done using a large dataset of sheep, n=174, from a single site, which presented a worst-case scenario given the high intra-site variabilities caused by a combination of anthropogenic and environmental influences.~~ Sample sizes larger than 40 presented minimal improvement to the true mean estimation (< 0.3 ‰). While sample sizes between 8–40 samples may prove ideal for estimating group means, it is important to remember that estimation of group ~~Approximation of group~~ variability may require larger ~~, on the other hand, seems to require larger~~ sample sizes ~~than estimation of group means~~ (Syväranta et al. 2013). Additionally, not all samples may pass quality control assessment and may thus yield unreliable results. For this reason, we recommend sampling slightly above the optimal number of samples (per species or group).

~~Power analysis can be a useful tool for determining the minimum number of samples required to answer a research question (Murphy et al. 2014). It requires estimates of likely effect sizes,~~

~~differences, and measures of variability (e.g. standard errors); these indicators can be obtained through pilot studies or using values from the literature. The information does not have to be precise, since the goal is to obtain a general idea of the sample sizes needed in order to detect, for example, differences of a specified magnitude. The analysis can be used to perform sample size calculations for a range of reasonable standard deviations, in order to determine how much the outcomes change as a result. This information can then help determine the optimal sample numbers that will be amenable for answering the question while avoiding redundancy. Important to keep in mind, however, is that not all samples may pass QC assessment, and may thus not yield reliable results. For this reason, sampling slightly above the desired number of samples of species per group is recommended.~~

Isotopic values of ~~F~~faunal remains can be used in palaeodietary reconstructions to estimate the underlying ecological 'baseline' values (discussed more in Section 2.6) or to investigate questions ~~sampled either to answer questions~~ pertaining to the management and dietary patterns of the animals themselves~~, to use in palaeodietary reconstructions as ecological 'anchors', or for both objectives combined~~ (e.g., Cucchi et al. 2016; Nitsch et al. 2017). As with sampling human remains, ~~The~~ faunal sampling strategies~~y for both of these aims should apply the same approach described above for sampling human remains, and~~should ~~carefully~~ consider ~~the sources and magnitures of variability, and~~ the benefit of carrying out power analyses that capture estimates of possible sources and magnitudes of variability. Since it cannot be assumed that animals from spatially or chronologically distinct horizons had the same diets, it is advisable to carry out test runs to assess the variability in the isotopic values of fauna from distinct contexts or levels~~ both aim to do an initial test run for assessment of intra-site or intra-period variability, and comprehensively sample distinct contexts where faunal remains are well preserved~~. If planning to reconstruct dietary food-webs using isotope mixing models (e.g. (Parnell et al. 2013; Fernandes 2016)~~use the results in Stable Isotope Mixing Models (SIMMS)~~, careful consideration should be given to identifying and quantifying possible ~~careful consideration needs to be given to the possible~~ sources of variability within the consumers, such as intra-individual variability in biological tissues and inter-individual differences in metabolic rates (Cheung and Szpak 2021; Hyland et al. 2021). Additionally, variability within the food items themselves (caused by, e.g., manuring, cooking and fermenting; Royer et al. 2017) and limitations posed by an incomplete

record of possible food items (where some may have not preserved in the archaeological record) should also inform sample selection and data interpretation. ~~(Royer et al. 2017) (e.g., intra-individual differences in biological tissues, inter-individual differences in metabolic rates(Cheung and Szpak 2021; Hyland et al. 2021){Citation}) within both the consumer and the food sources~~For detailed advice on sampling and using dietary mixing models, ~~and~~ we suggest that researchers follow the best ~~-~~practices outlined in Cheung and Szpak (2021).

~~However, u~~Unequal representation of ~~different~~ distinct ~~skeletal~~ elements ~~within~~ in ~~the faunal~~skeletal assemblages often poses a limitation to identifying the desired Minimum Number of Individuals (MNI) from each archaeological context of interest. When ~~the bone~~ skeletal assemblages are~~is~~ large (~~as, for~~ e.g., the ~~animal bone~~faunal assemblage at Neolithic Makriyalos, northern Greece), the issue of sampling unique individuals more than once can be circumvented by selecting bones of a single element and body side for each taxon~~, taking into account the side of body and size~~. At Makriyalos, Paul Halstead and his zooarchaeological team determined that selecting *humeri* from domestic herbivores and *radii/tibiae* from red deer would provide ~~them with~~ the largest ~~Minimum Number of Individuals~~MNI for each taxon (Vaiglova et al. 2018). At sites with smaller/less well-preserved bone assemblages (e.g., Kouphovouno, Vaiglova et al. 2020a), it may not be possible to only sample one element. In this instance, ~~and~~ the researchers should ~~thus~~ make the best effort to minimize the possibility of sampling an individual more than once~~,~~ while achieving the ~~context~~ coverage necessary to address the research question. In both scenarios, including information about the samples (element, side, element portion, and other relevant information, e.g., ~~(~~Balasse et al. 2013) can help demonstrate the researchers' confidence in having sampled distinct individuals.

Preservation of bones is determined by the ability of the~~its~~ organic and inorganic phases to survive in distinct climatic and pedogenic contexts. The organic phase is primarily composed of collagen, a protein that can remain intact in archaeological contexts up to 100,000 years old (Jones et al. 2001), unless ~~and unless~~ an archaeological site is located in the tropics, or in regions with very hot/dry summers and wet winters~~, it can remain intact in archaeological contexts up to 100,000 years old (Jones et al. 2001)~~. Because bone collagen enables measurement of both stable carbon and nitrogen isotope values, ~~It is for this reason that bone collagen~~it has become the preferred

sampling material for palaeodietary studies~~,~~; although it primarily provides information about the protein component of consumers' diets (Sullivan and Krueger 1981; Ambrose and Norr 1993; Lee-Thorp 2008).

With its high porosity and low crystallinity, ~~T~~the inorganic bone mineral bioapatite ~~has high porosity and low crystallinity compared to the robust enamel bioapatite (see more below), and~~ can become significantly altered through long-term interaction with soil (Lee-Thorp 2008; Clementz 2012). The surface area of bone ~~crystalline structure~~ is amenable to bicarbonate adsorption (Koch et al. 1997), which can likely alter~~ing~~ the stable carbon ($\delta^{13}C$) and oxygen ($\delta^{18}O$) isotope composition of bone apatite. Nevertheless, ~~Non-calcined b~~one bioapatite can provide reliable stable isotopic measurements ~~under certain preservation conditions —~~when it follows "a pathway to preservation, rather than to destruction" (Lee-Thorp and Sponheimer 2003, p. 210). This can happen in dry burial environments which~~that~~ favor slow removal of collagen with relatively fast increase in crystallinity ~~, which can~~ leading to million-year-old bone resembling enamel in its crystallographic features (Lee-Thorp and Sponheimer 2003). Thus, preservational context plays a role in the potential of bone mineral to retain its *in vivo* stable isotopic composition and should be carefully considered when identifying bones suitable for isotopic analysis.

~~Similarly, s~~Strontium, a mobile trace element in soils and sediment, can easily substitute for calcium in bone bioapatite to the point where it ~~i~~becomes~~s~~ impossible to distinguish between *in vivo* strontium and contamination from soil. ~~determine whether strontium present in the bone sample represents *in vivo* strontium or contamination.~~ Experiments have shown that bones that are calcined (i.e. heated to ~~cremated at~~ temperatures that turn them white) can preserve their *in vivo* $^{87}Sr/^{86}Sr$ ~~$^{87}Sr/^{86}Sr$~~ strontium isotope ratios~~ratios~~, $^{87}Sr/^{86}Sr$, (Snoeck et al. 2015), but not their $\delta^{13}C$ or $\delta^{18}O$ values (Snoeck et al. 2016). For this reason, while calcined bone may provide reliable samples for strontium isotope ratio measurements in studies of residential mobility, non-calcined bone should be avoided or used only to determine local baseline $^{87}Sr/^{86}Sr$ ratios. ~~Non-calcined bone bioapatite can provide reliable isotopic measurements under certain preservation conditions —when it follows "a pathway to preservation, rather than to destruction" (Lee-Thorp and Sponheimer 2003, p. 210). This can happen in dry burial environments that favor slow removal of collagen with relatively fast increase in crystallinity, which can lead to million-year-old bone resembling enamel~~

~~in its crystallographic features (Lee-Thorp and Sponheimer 2003). Thus, preservational context plays a role in the potential of bone mineral to retain its *in vivo* isotopic composition and should be carefully considered when identifying bones suitable for isotopic analysis.~~

2.2 T~~ooth enamel~~eeth

Tooth enamel is commonly analyzed using varying scales of resolution. 'Bulk' tooth samples (i.e. one measurement per individual; e.g., Roberts et al. 2015) approximate the average value of all isotopic inputs over the course of tooth formation. This can provide a way to compare multi-seasonal averages between individuals. Similarly to bone collagen sampling discussed above, ~~Teeth can be sampled either in bulk (i.e. one measurement per individual, e.g. Roberts et al. 2015), or incrementally using a series of sub-samples drilled along the axis of the tooth's growth (Zazzo et al. 2006; Balasse et al. 2013). Bulk sampling of individuals can proceed in the same way as sampling of bones for bulk collagen analysis~~power analyses can be useful for estimating the optimal number of samples required to answer the research question~~, with research aims and power analyses determining how comprehensively to sample different archaeological contexts in order to approximate the site-wide means/variability, or the context-specific means/variability~~ (Pearson and Grove 2013; Syväranta et al. 2013; Cheung and Szpak 2021).~~;~~ ~~with initial test runs providing helpful insight to help guide optimal sampling.~~

Incremental sampling that follows the direction of tooth growth provides ~~of enamel along the axis of growth provides~~ a sequence of samples that can be used to assess sub-annual changes in isotopic inputs throughout tooth formation (Balasse 2003). The temporal resolution of these sequences varies depending on the sampling technique. Manually sampled sequences (~~targeting~~extracting 1mm–wide samples spaced ~1mm apart, e.g., Balasse ~~(~~2012; Janzen et al. 2020) provide attenuated values averaged over several months of tooth formation (Balasse 2003; Zazzo et al. 2005; Pederzani and Britton 2019). Micro-driling with laser-ablation mechanisms (e.g., ~~(~~Copeland et al. 2008; Pryor et al. 2020) provide higher resolution time slices averaged over a few months, while targeted micro-drilling using histological maps of developmental features (i.e. daily growth lines) can approximate isotopic inputs with weekly resolution (e.g., ~~(~~Austin et al. 2016; Smith et al. 2018).

Due to budgetary constraints, incremental sampling ~~is often~~often necessitates analysis of fewer individuals than bulk sampling. ~~more restrictive to MNI than bulk sampling.~~ For example, while Roberts et al. (2015) were able to analyze 300 individuals with bulk sampling, incremental sampling was more restrictive: ~~Balasse et al.~~Janzen et al. (2020) analyzed 60~~15~~ individuals, Pryor et al. (2020) analyzed 16 individuals and Smith et al. (2018) analyzed 8 individuals. ~~Makarewicz et al. (2017) analyzed 15 individuals, and Janzen et al. (2020) analyzed 30 cattle and 30 caprine individuals.~~ For this reason, incremental tooth enamel sampling is ~~usually used to~~most amenable for answering questions of individual or group, rather than population-wide, life histories. ~~individual or group specific, rather than population-wide, research questions.~~

A major consideration of sample selection is the likelihood that a tooth preserves a ~~full~~ complete annual sequence representative of the full range of variability in seasonal isotopic inputs. Partial sequences from highly worn teeth may represent incomplete annual cycles, and the absence of 'real' maximum or minimum values may preclude the ability to determine which seasons are represented. For example, of the 32 teeth drilled from Byzantine sites in the Negev desert (Vaiglova et al. 2020b), 10 teeth provided incomplete sequences, which are shown in Supplementary Material 2. Without being able to determine which seasons are represented, inclusion of these partial sequences in the analyses would lead to skewed results (e.g., over-representation of summer values and under-representation of winter values). To overcome this problem, only teeth with full annual sequences should be included in ~~the formal analyses~~inter-teeth comparisons ~~(e.g. Vaiglova et al. 2020b drilled and measured incremental samples from 32 teeth, but only 22 provided full annual sequences were included in the analysis)~~. If it is possible to determine which seasons are represented (e.g., by identifying a clear sinusoidal 'peak' or 'trough'), partial sequences may be used to compare seasonal values (e.g., compare only summer values across individuals). Teeth that complete mineralization over several annual cycles (e.g., horse teeth, ~~Teeth that take significantly longer than an annual cycle to complete mineralization (e.g. horse teeth with mineralization schedules spanning ~1.5 to 2.8 years,~~ Hoppe et al. 2004: Table 2) may also provide skewed results if teeth of varying wear stages record formation over varying numbers of annual cycles. ~~distinct numbers of seasons are represented across different teeth.~~ To overcome this issue, fixed~~pre-determined~~ lengths of teeth crowns (e.g., 3cm for horse teeth, Britton et al. 2019) can be drilled to keep the temporal resolution as constant as possible.~~approximate the~~

~~dietary inputs over a single annual cycle (e.g. Britton et al. (2019) obtained incremental samples from 3-cm regions of horse third molars).~~

The choice of tooth is constrained by the timing of tooth mineralization and the ability of different teeth to answer the research question. Unlike bones, which remodel during the lifetime of individuals, the elemental composition of teeth remains ~~overwhelmingly~~ generally constant once formation is complete (Pederzani and Britton 2019). The precise timing of tooth development is primarily determined by species and tooth type., ~~but e~~Earlier forming teeth (e.g., first molars, M1) can provide information about initial life patterns such as cessation of nursing, while later forming teeth (e.g., third molars, M3) record dietary inputs in later years (Weinreb 1964; Balasse et al. 2001; Hillson 2005). The duration of mineralization also determines if the tooth will record full annual sequences or not. For example, due to their shorter lengths and faster mineralization schedules, pig molars do not record full annual sequences and are thus not the tooth of choice for studying ~~because they are shorter in length and faster forming than herbivore molars, pig molars do not record a full annual sequence of dietary intake, and assessment of their~~pig seasonal dietary and mobility patterns. These objectives can instead be accomplished by incrementally sampling their longer-forming incisors (Frémondeau et al. 2012; Zhang et al. 2021).

Unlike bone apatite, tooth enamel is relatively robust to diagenesis (Zazzo et al. 2004; Lee-Thorp 2008). Its higher crystallinity prevents exogenous trace elements like strontium from being incorporated deep inside the mineral structure (Macfadden and Cerling 1996; Snoeck et al. 2015). More recent research, however, has revealed that tooth enamel is not always as structurally robust as we routinely assume (Wood et al. 2021; Weber et al. 2021). ~~It is advisable to apply QC~~ Quality criteria outlined in Table 1 can help ~~to~~ assess the integrity of tooth enamel for isotopic analysis prior to analysis of the results. For high~~er~~-resolution analyses that employ instruments such as a ~~including~~ Laser Ablation Inductively Coupled Plasma Mass Spectrometer~~s~~ (~~LA-ICP-MS;~~ e.g., Sponheimer et al. 2006), or Secondary High Resolution Ion Microprobe (~~SHRIMP;~~ e.g., Aubert et al. 2012), elemental mapping of trace elements like uranium can be used to identify portions of enamel that may have undergone diagenetic alteration. This information can help ~~and direct analysis to~~target regions that are most likely to provide *in vivo* composition.

Depending on the questions, researchers working with tooth enamel may choose to analyze either $\delta^{13}C$ and $\delta^{18}O$ values of structural carbonate ~~of tooth enamel bioapatite (for measurement of $\delta^{13}C$ and $\delta^{18}O$ values)~~ or $\delta^{18}O$ values of ~~the enamel~~ enamel phosphate ~~(for measurement of $\delta^{18}O$ values)~~. Enamel ~~The more structurally robust~~ phosphate is more resistant to diagenetic alteration (Zazzo et al. 2004), and is the preferred material for addressing questions of palaeoclimate~~.~~. On the other hand, matching $\delta^{18}O$ and $\delta^{13}C$ values of structural carbonate makes these samples amenable to addressing questions of both seasonal mobility and dietary intake ~~On the other hand, the ability to measure both $\delta^{13}C$ and $\delta^{18}O$ values makes carbonate useful for answering questions about mobility as well as dietary intake~~ (Makarewicz and Pederzani 2017; Pederzani and Britton 2019).

2.3 Charred plant remains

Sample selection for plant ~~stable~~ isotope analysis is also constrained by both the research question and the depositional history of the assemblage. ~~Cereal grains~~Seeds from primary *in situ* contexts (e.g., storage deposits) can provide information about growing conditions on ~~a~~ single years~~/~~ or locations (e.g., Styring et al. 2016)~~,~~. S~~while grains~~ eeds derived from secondary contexts necessarily provide an averaged indicator of cultivation conditions that can span multiple years or ~~/~~locations (e.g., Vaiglova et al. 2020a). ~~Grain~~ Seed samples can either be obtained in bulk (i.e. several ~~grains~~ seeds from an individual archaeological context) or as individual single~~-grain~~ seed samples, and the choice depends on the research aims. Single seed analysis can be useful for assessing inter~~-plant~~crop variability within sampling contexts, and to avoid the possibility of averaging crop growing conditions across different cultivation events. However, this analysis may not be feasible in cases where the amount of nitrogen (%N) of small samples is too low for reliable analysis of stable nitrogen isotope values ($\delta^{15}N$) ~~measurement~~ (Szpak and Chiou 2019).

Formatted: Font: Not Italic

Bulk sampling is effective if the ~~grains~~ seeds derive from primary contexts where it can be assumed that the crops represented were cultivated in the same year~~-~~; and ~~(~~this only applies to species of annual plants~~)~~. It can also be an appropriate sampling strategy if the assessment of inter-plant variability is not ~~of concern for~~crucial answering the research question. Experimental research on modern crops from a single growing season and treatment condition suggests that analysis of homogenized bulk samples of ~10 cereal grains register variability of ±0.5 ‰ in $\delta^{13}C$ values and ± 1.0 ‰ in $\delta^{15}N$ values ~~with 95% confidence intervals~~, while analysis of smaller numbers of grains

(i.e., < 5) show much greater variability (Nitsch et al. 2015). Therefore, to assess the variability between growing conditions represented in single *in situ* deposits or storage contexts, we recommend that researchers ~~sampling should~~ aim to obtain several (ideally more than 3) bulk samples per context.

The most common way seeds are preserved in the archaeological record is via charring, although uncharred or lightly toasted seeds have been found in rare waterlogged, frozen, or desiccated contexts. Outside of these rare preservation conditions, seeds need to be exposed to temperatures above ~215 ºC to become highly resistant to microbial attack and survive long-term burial. Seeds charred above ~400 ºC become highly distorted and unamenable to species identification (Charles et al. 2015). Because of their highly porous structure, charred material can absorb high amounts of strontium from soil and reliable application of $^{87}Sr/^{86}Sr$ ratio analysis for provenancing archaeological plant material is currently highly problematic (Styring et al. 2019).

For analysis of stable carbon and nitrogen isotope values, selection of ~~Selection of~~ charred ~~grain~~seeds ~~samples also~~ requires the ability to assess the effects that charring has had on the *in vivo* isotopic composition of the material. Experimental studies have shown that temperature, and to a lesser degree duration, alter both $\delta^{13}C$ and $\delta^{15}N$ values of cereal grains and pulse seeds (Nitsch et al. 2015; Stroud et al. in prep). These experiments have ~~further~~ also shown that the isotopic values of seeds charred between 215–300 º~~C~~elsius are altered in ways that can be modelled~~,~~ and corrected ~~allowing for such alteration to be accounted for with a charring correction~~ (Nitsch et al. 2015; Stroud et al. in prep). Ascertaining the temperature at which seeds have been charred requires visual inspection of both the surface morphology and the internal ~~charred matrix~~seed structure (cf Charles et al. 2015, Stroud et al. in prep). Assessment of the internal matrix is crucial, as recent experiments have shown that even cereal grains whose surface morphological characteristics suggest that they are 'ideally' preserved ~~that look 'ideally' preserved externally~~ can be completely void on the inside (Stroud et al. in prep). Because the preservation potential of cereal grains can vary within and between archaeological contexts, plant material should be assessed on a grain-by-grain basis to ensure that only ~~seeds~~ grains that are sufficiently preserved are measured and interpreted (Stroud et al. in prep). To demonstrate confidence in sample reliability, we suggest that researchers publishing ~~To publish~~ plant stable isotope results~~, we~~

~~suggest that researchers~~ provide access (in supplementary materials or online databases) to photographs of each analyzed seed that is included in the analysis~~include photographic evidence of the seed cross-sections to demonstrate confidence in sample reliability~~.

~~Additionally, t~~The isotopic and morphological changes caused by ~~that~~ charring ~~produces~~ are currently only understood for a relatively small number ~~range~~ of cereals: emmer (*Triticum dicoccum* (Schrank) Schübl.), einkorn (*Triticum monococcum* L.), hulled barley (*Hordeum vulgare var distichum* L.), bread wheat (*Triticum aestivum* L.), spelt (*Triticum spelta* L.), oat (*Avena sativa* L.), ~~and~~ rye (*Secale cereale* L.), maize (*Zea Mays* L.), and one millet species: pearl millet (*Pennisetum glaucum* L.); and two pulse species: lentil (*Lens culinaris* Medik.) and pea (*Pisum sativum* L.)~~, and one millet species: pearl millet (*Pennisetum glaucum* L.)~~ (Nitsch et al. 2015; Charles et al. 2015; Styring et al. 2019; Hart and Feranec 2020; Stroud et al. in prep)~~(Charles et al. 2015, Nitsch et al 2015, Styring et al 201(Hart and Feranec 2020)9, Stroud et al. in prep)~~. For this reason, before further charring experiments are carried out, caution should be taken when interpreting isotopic data derived from species for which the effect of charring is unknown.

2.4 Molluscan shell and eggshell

Shells of aquatic and ~~marine and~~ terrestrial molluscs preserve valuable short-term and incremental records of palaeoenvironmental parameters such as water temperature and C_3/C_4 plant abundance (Goodfriend 1992; Johnson et al. 1998; Andrus 2011). However, aspects of the animal's life history, habitat, and physiology may incorporate systematic biases and offsets into the preserved environmental record (Schöne 2008; Ivany 2012). An animal typically slows down or speeds up shell production under particular environmental conditions (e.g., temperature, food availability) or physiological circumstances (e.g., reproductive cycle, recovery from predatory attacks), resulting in those periods becoming either over- or underrepresented in their internal record. Additionally, an animal's behavior may change over the course of its life. For example, many marine molluscs will move to higher or lower regions within the intertidal zone as they grow larger (and thus record a changing temperature range across their shell), or their ~~annual~~ growth rate will slow down with age (Schöne 2008). As these factors are unique to each species, it is important to carefully consider the ecology and biology of the given species. A~~, and undertake a~~ robust study of shells from modern and well-characterized ~~(i.e. modern)~~ environments can help ~~in order to~~ to assess the

sampling resolution required to attain the desired temporal resolution of the temperature record (e.g., daily, monthly, annual), as well as to evaluate any systematic biases (e.g., seasonal, ontogenetic). It can also provide information on any inter-individual variation caused by differences in life histories and microhabitats, which should be taken into consideration when deciding how to apportion samples within individuals or entire assemblages (e.g., (Galimberti et al. 2017). and applicability of their stable isotope measurements for reconstructing annual and/or sub-annual conditions.

In addition to validating the usefulness of particular shell-producing species for stable isotope studies, shell samples need to be assessed for preservation (Leng and Lewis 2014). Carbonate shells are prone to dissolution and recrystallisation in many depositional settings where water is present in the sediments. Unfortunately, the process of recrystallisation may overwrite the original environmental signal as recorded in their $\delta^{18}O$ and $\delta^{13}C$ values, and such these shells should be regarded as unsuitable for stable isotope analysis (Brand 1989). Additionally, cooking practices may have also altered the shell structure, and such that burned or heated shells must likewise be avoided (Milano et al. 2016). For these reasons, it is necessary to evaluate the crystal microstructure of individual shells, especially where preservation varies considerably across a site.

Visual examination of sectioned shells under a microscope can quickly reveal changes to shell structure, with recrystallized shell looking amorphous and 'flaky', rather than highly crystalline. Many aquatic species and almost all terrestrial species precipitate part or all of their shell in the aragonitic polymorph of carbonate, in the shell material from highly crystalline to amorphous and 'flaky'. Amongst some marine species and almost all terrestrial species, all or part of the shell is precipitated in aragonite as opposed to calcite (Lowenstam 1954; Goodfriend 1992). A This offers a convenient diagenetic check as recrystallized shell is always calcitic, and the two polymorphs recrystallized carbonate will be calcitic; and calcite and aragonite are simple to distinguish with a range of analytical chemistry methods (i.e. XRD, Raman spectroscopy, chemical staining), assessment of the presence of aragonite and calcite offers a convenient diagenetic check during sample selection. , although some are more destructive and/or less quantitative than others.The methods, however, vary considerably in terms of how destructive and/or quantitative they are. FTIR (Fourier Transform Infrared Spectrometry) with an ATR (Attenuated Total Reflectance)

attachment offers a suitable balance of analytical speed and simplicity, ~~sample~~ damage to the sample, and polymorph ~~and recoverability, and calcite/aragonite~~ quantification (Loftus et al. 2015).~~,~~ This enables each powdered sample to be assessed for diagenesis prior to stable isotope analysis. ~~allowing each analytical powder sample to be assessed for diagenesis prior to stable isotope analysis.~~

2.5 Samples obtained during AMS radiocarbon dating

In radiocarbon dating laboratories, $\delta^{13}C$ values are measured alongside radiocarbon determinations for several reasons: 1) ~~in order~~ to correct for reservoir effects which may have shifted the measured age away from the true age (e.g.~~,~~ marine reservoir effect, ~~(~~Knipper et al. 2020), 2) ~~and~~ to correct for fractionation to the ^{14}C:^{12}C ratio that occurred during the graphitization stage of sample preparation, and 3) to systematically normalize all radiocarbon measurements to $\delta^{13}C = –25.0$ ‰, even if they have not been influenced by reservoir or graphitization fractionation effects (Stuiver and Polach 1977; Reimer et al. 2004). These measurements are performed either directly on the Accelerator Mass Spectrometer (AMS) or by directing a portion of the CO_2 gas from the AMS to the Isotope Ratio Mass Spectrometer (IRMS). However, these measurements are often not usable for palaeodietary/palaeoclimatic interpretations, or for identifying small differences between samples (e.g.~~,~~ < 0.2 ‰) for three reasons (personal communication: Richard Staff, Scottish Universities Environmental Research Centre; David Chivall and Peter Ditchfield, Oxford Radiocarbon Accelerator Unit).

First, the graphitization stage during sample preparation induces fractionation, which essentially alters the ^{13}C:^{12}C ratio by an indeterminate amount. Secondly, much larger samples are injected into the AMS than would be suitable for IRMS and this can push the IRMS to its limit, potentially introducing error. Third, the AMS $\delta^{13}C$ values are often normalized using a one-point rather than two- or multi-point calibration, which leads to lower measurement accuracy. At the Oxford Radiocarbon Accelerator Unit, ~~T~~the uncertainties attached to AMS $\delta^{13}C$ measurements are not calculated, but ~~at the Oxford Radiocarbon Accelerator Unit,~~ the system flags measurements only when the $\delta^{13}C$ value obtained during combustion for graphitization and the $\delta^{13}C$ value obtained by the AMS differ by > 5.0 ‰. For this reason, when clients request accompanying $\delta^{13}C$ values, separate aliquots of the samples are prepared for analysis using an IRMS, where the uncertainty is

often better than 0.5 ‰. Altogether, stable carbon isotope values produced via AMS dating are not the same as those obtained on an IRMS. ~~and r~~Researchers should thus only use AMS $\delta^{13}C$ values for palaeodietary/palaeoclimatic research when their accuracy and precision can be determined and when the measurement error fits within the expected uncertainty for answering the research question. ~~values were calibrated using multi-point calibration and when the measurement uncertainties make it feasible to address the research question.~~

2.6 Final word on sampling: disentangling natural from anthropogenic factors

The extent to which isotopic measurements of archaeological materials can shed light on questions of ~~past dietary habits, mobility patterns, manipulation of landscapes and plant/animal resources, or shifting climatic patterns~~archaeological interest is constrained by our ability to isolate the effects of human ~~choices~~ behavior from the effects of environmental factors that may have simultaneously influenced the chemical composition of the sample material. ~~concomitantly influence the chemical composition of these materials.~~ For interpreting strontium isotope ratios, comparative maps of biologically available strontium can be built by modelling the spatial distributions of $^{87}Sr/^{86}Sr$ ratios of modern plants, surface water, snail shells, insects (Evans et al. 2010; Snoeck et al. 2020; Wang et al. 2021; Holt et al. 2021). For interpretation of archaeological stable carbon and nitrogen isotope values, the ecological background can be approximated through measurement of ~~This can be attempted by obtaining and measuring~~ specimens with ~~reduced~~ low likelihood of ~~-~~impact from anthropogenic ~~influences~~ factors (e.g., wild species, unmanaged landscapes). However, the ~~identification of which samples would provide~~identification of non-anthropogenically influenced samples is fraught with many challenges.

Firstly, rather than focusing on estimating average base*line*, we should instead focus on approximating a range (or 'base*interval*') of local background values. Baseintervals will be able to capture both the natural variability within the given ecology and the measurement error accompanying the isotopic measurement. ~~robust analogues for building a 'baseline' (or more appropriately 'base*interval*', as the underlying values will undoubtedly exhibit natural variation and their estimation will be accompanied by measurement error), is fraught with many challenges.~~

Secondly, baseintervals should not be thought of as static. Since isotope ecologies are determined by a range dynamic factors including ~~Firstly, the underlying 'baseinterval' values may have changed on seasonal, annual, decadal, or millennial scales as a result of fluctuation in~~ precipitation patterns, soil dynamics, and ~~shifts in the~~ distribution of ~~composition of~~ local vegetation and faunal species, they may have changed on annual, decadal, centennial, or millennial scales ~~distribution~~ (Gannes et al. 1997). These changes would have caused reverberating effects through the food chain and ~~,~~influenced~~ing~~ the isotopic composition of ~~both the~~ food sources (primary and secondary producers)~~,~~ as well as consumers ~~(belonging to~~in ~~different~~ distinct trophic levels~~)~~. Measurements of reference samples that are as contemporary as possible to the samples of interest are ~~*contemporary* reference samples is~~ therefore key for demonstrating that observed patterns in isotopic data are most likely to be the ~~a~~ result of changes in human behavior ~~and choices~~ rather than shifts in the underlying ecology.

~~Secondly~~Thirdly, multiple ~~floral~~ plant and ~~faunal~~ animal species may be necessary to capture the isotopic variability of specific vegetation communities, ungulate dietary niches, or human food production systems (Codron et al. 2011, 2013). This is where questions ~~However, questions~~ may arise regarding whether or not the comparative species may have been ~~were~~ impacted by anthropogenic ~~influences themselves~~factors, especially in archaeological contexts where intensive agriculture or landscape management was pursued. ~~for surplus production and provisioning of urban places was pursued.~~ For example, wild herbivores may have raided crop fields, and their bone collagen δ^{13}C and δ^{15}N values ~~and their δ^{13}C and δ^{15}N values~~ would thus reflect the consumption of both natural forage ~~as well as~~and plants that may have been influenced by manuring, irrigation or crop selection (e.g.~~,~~ ~~^{13}C enriched millet or maize in temperate C_3 biomes or ^{13}C depleted barley or wheat in a C_3/C_4 steppe biome,~~ Gillis et al. 2020). ~~In addition, h~~Human activities such as woodland clearance ~~for collection of timber or creation of pastures~~ may have forced wild herbivores to graze in open environments with higher plant δ^{13}C values compared to those in closed–canopy woodlands ~~to graze in more open environments, and subsist on plants with higher δ^{13}C values due to reversal of the 'canopy effect'~~ (Bonafini et al. 2013). ~~Similarly~~Additionally, ~~wild plant species are not necessarily likely to be free of anthropogenic inputs, as~~ weeds may have grown in arable cultivation plots~~, and~~ while nut or fruit trees may have

been planted on the edges of managed fields, where they may have been similarly influenced by human activities that impact the $\delta^{13}C$ and $\delta^{15}N$ composition of soil. ~~.~~

~~In order to~~To identify faunal species most appropriate for building reference ~~'~~baseintervals~~'~~, it is important to consider the specific dietary adaptations (e.g., browsing vs. grazing) and physiological requirements (e.g., obligatory vs. non-obligatory drinking) of the available species. ~~the specific dietary adaptations and physiological requirements of the species available from a given faunal assemblage while also considering how adaptive responses and requirements of ancient species may have differed from those of modern-day taxa.~~ Additional considerations include ~~ly,~~ questions of population size, forage availability, and predation pressure ~~need to be borne in mind~~ , as these ~~they~~ may have influenced dietary niche size and composition and~~,~~ caused~~ing~~ directional isotopic changes to wild animal populations over time (Codron et al. 2011).

Identifying comparative plant species needs to take into account both the likelihood that they grew in managed soils, as well as possible isotopic variability between distinct plant parts and growing seasons ~~Comparative plant species need to be assessed in terms which functional groups they belong to, what likely isotopic variability exists between distinct plant parts~~ (Tieszen 1991; Fraser et al. 2011; Cernusak et al. 2013). Although the directional effects of certain crop management systems (e.g. application of farmyard manure or other organic waste, small or large-scale systems of water management) have been demonstrated experimentally ~~While experimental studies have shed light on the directionality of anthropogenic activities such as soil or water management (including more or less formalized systems of irrigation, hand watering, and application of organic waste or farmyard manure on soils to increase their productivity;~~ (Bogaard et al. 2007; Fraser et al. 2011; Wallace et al. 2013; Flohr et al. 2019; Jones et al. 2021), the magnitude of the effects would be challenging to estimate in environments where experiments have not been carried out. ~~understanding the same effects in different environments may prove challenging.~~ For example, using the manuring offset of crops that were grown ~~magnitude of the manuring effect observed in crops that were experimentally grown~~ in Britain, Germany, and Romania (Fraser et al. 2011) to estimate the degree of manure application on crops grown in South Africa ~~disentangle between the magnitude of natural and anthropogenic ^{15}N enrichment of soils and plants in South Africa~~ would

be inappropriate, because the two environments are not directly comparable due to differing climate, soil types, and preservation potential of archaeological materials.

With these caveats in mind, sampling for comparative baseintervals should aim to capture, to the best of ~~our abilities~~a researcher's ability, the full range of isotopic variability within a ~~encompassed by~~ given ecosystem~~s (whether they be canopied forests, open grasslands, wetland)~~. Interpretations of measured archaeological values should then be constrained by~~, and constraining the interpretations within~~ predictions ~~made about~~surrounding the extent to which the measured species may have been ~~their forage was~~ influenced by natural ~~vs.~~and anthropogenic factors (Roberts et al. 2017b; Knipper et al. 2020; Gillis et al. 2021). ~~If possible, sampling faunal or floral species from cultural-chronological periods outside of those that represent the core of the study may help distinguish between environmental and anthropogenic factors that may have influenced the individuals of interest. Furhtermore~~Additional~~, other~~ lines of evidence, such as pollen records, ~~inferences of~~ patterns of agro-pastoral intensification and ~~and~~ extensification ~~of agro-pastoral activities~~ (Styring et al. 2017; Price et al. 2020; Marston 2021), human landscape modification (Rull et al. 2021), and local-scale environmental change (Roffet-Salque et al. 2018) can furthermore be used to ~~strategically integrated with the comparative baseintervals to further~~ clarify the mechanisms that may have influenced the isotope composition of archaeological flora, fauna, and human remains.

~~Effective sampling for building comparative baseintervals should thus focus on accomplishing three goals:~~

~~1) Understanding isotopic variability across multiple taxonomic groups to document varying, but relevant, ecological niches or aspects of food production systems.~~

~~2) Obtaining large enough samples to account for the range of isotopic variation in plant and animal communities that interacted with humans.~~

~~3) If change over time is of research interest, covering the relevant successive chronological phases and (if possible) periods that preceed or succeed the phases of interest to build a more robust comparative framework.~~

3. Using exploratory data analysis to guide formal analysis

Exploratory data analysis ~~(EDA), or more generally the graphical display of data,~~ is an important part of any analysis and should be carried out prior to formal analysis through inspection of graphical representation of the data. ~~-~~Simple univariate plots such as histograms, boxplots, and violin plots can shed light on the behavior of individual variables, while scatterplots can elucidate the connections between them. ~~EDA~~Exploratory data analysis should not be a rote exercise, but ~~rather~~ should instead involve a thoughtful investigation of the data ~~(not necessarily using the default settings in software)~~. Many books have been written about the principles of good data visualization (e.g., Cleveland 1994; Wilkinson 2005) and packages such as ggplot2 in R (Wickham 2016) provide a variety of flexible tools for effectively exploring data.

3.1 Suggestions for effective data visualization

Simple aesthetic adjustments and choices can have a large ~~effect~~ impact on the readability of figures. One does not need to be proficient with coding~~,~~ nor to use programs like R and Python to produce graphs that are aesthetically pleasing and easily readable. Here we outline some of the most relevant advice for visualization of isotopic data; referenced examples are reproduced in Supplementary ~~Materials~~ Material (SM) 1~~(SM)~~.

(a) ~~(a)~~ Plot the data in ways appropriate for addressing the research question. If the objective is to assess how much variation there is between individual data points, standard deviation is most effective for showing the spread of the data. If the objective is to assess how good the estimation of the group mean is, standard error is more appropriate. Boxplots, either individual or plotted side by side to show group comparisons, can highlight important and interesting ~~behavior~~patterns in datasets. For archaeological data~~,~~ which may be non-normally distributed (bimodal, skewed), violin plots may be more informative than boxplots (e.g., Cavazzuti et al. 2019, Fig.6, in ~~S~~SM~~M~~ 1.A). Bivariate scatterplots (e.g., Levin et al. 2011, Fig.8, in SM~~M~~ 1.B), especially when enhanced by scatterplot smoothers (such as loess curves, e.g., Wierzbowski et al. 2017, Fig.8, in S~~M~~M 1.C), are an effective way of exploring the relationships among variables~~)~~. -The idea can, with care, be extended to rotating three-dimensional plots that may reveal relationships that are not evident from collections of bivariate plots alone (see, for example, Wainer 2007, pp. 126–127).

Crucially, the most appropriate method of visualization will depend on the relationships being explored. For example, with intra-tooth sequences of isotopic values, one can look at the *shape* of the sequences to gather information about seasonal variability in dietary inputs (e.g., Vaiglova et al. 2018, Fig.10, in ~~SM~~ SM 1.D), the *intra-tooth variability* to assess the magnitude of variation in dietary inputs between different individuals (e.g., Vaiglova et al. 2020b, Fig.2, in S~~M~~M 1.E), or the *position* of different parameters (such as the maximum $\delta^{18}O$ value) on the tooth axis (e.g., Balasse et al. 2021, Fig.2, in S~~M~~M 1.F). Figures can thus be used appropriately to highlight these relationships of interest. See ~~also~~ Weissgerber et al. (2019) for advice and examples of matching visualization to research question.

(b) Do not use randomly assigned combinations of symbols and colors. ~~Look~~ Looking at the spread of data to see ~~at the data to see~~ which species, groups, or individuals overlap allows researchers to ~~and~~ choose symbols based on what will make the data points most easily discernible. The size of plotting symbols can often be controlled (~~to be~~ made smaller or larger), while color can be shaded and made more or less transparent. This can be used to the researcher's advantage, yielding more flexibility beyond the software defaults. For example, in the dataset presented in Vaiglova et al. (2021, Fig.S2, in S~~M~~M 1.~~F~~G), many human datapoints overlapped with those of dogs and pigs. Thus, the authors chose to use different colors, rather than symbols, to represent these three species. Where less overlap was visible (between humans and cattle/sheep/goats), different symbols were used. Altogether, only 5 colors were used to represent 18 species, and the color-symbol combinations were used in ways that led to least overlap between similar symbols.

(c) Keep graphs as simple as possible. Overcrowding figures with too many overlapping groups may obscure actual patterns of interest from being discernible. For example, if individuals from distinct age or sex categories have meaningfully different isotopic values, it is useful to display them separately. However, if their values are not meaningfully different, they can be combined into fewer groups to prevent the reader's eye from trying to detect patterns where they do not exist. ~~Where groups do overlap, it may be more effective to display them together and comment on this in the text. A common example of this is displaying categories such as "females?", "males?", "indeterminate age". If these individuals stand out statistically, a point can be made about displaying them separately, but if they do not, it is advisable to combine them into fewer groups~~

~~to avoid having the reader's eye try to detect patterning where it does not exist.~~ Where a large number of units are chosen to be displayed, an effective strategy is to group them into clusters that share the same symbol, and that are further differentiated with different colors (e.g.~~,~~ Knipper et al. 2017, Fig.3, in S~~M~~M 1.H~~G~~)

(d) Make graphs readable to people with color vision deficiency. Millions of people around the world suffer from color vision deficiency (CVD), with prevalence for just one type of deficiency (red-green) reported to be between 4–8% in men and ~0.4% in women (Birch 2012). To make publication figures as readable as possible to the wider audience, we suggest that researchers use CVD-friendly palettes to prepare their graphical images, and run them through a simulator (e.g.~~,~~ https://www.color-blindness.com/coblis-color-blindness-simulator/) to check that they are compatible with different types of deficiencies. If CVD-friendly palettes are not used, using black borders around symbols (as opposed to colored borders) makes the colors more discernible. The combination of green and red is most commonly challenging, but different shades of purples and oranges can work well with the full range of color deficiencies. Vaiglova et al. (2021, in SM~~M~~ 1.G~~F~~) used the IBM palette to make patterns in a scatter plot of $\delta^{13}C$ vs. $\delta^{15}N$ values visible on a yellow-shaded background. If combinations like green and red are used, the groups can be distinguished using different symbols (e.g.~~,~~ Levin et al. 2011, Fig.7, in S~~M~~M 1.B), but this only works with small numbers of groups.

(e) Use legends to your advantage. ~~Although m~~Many researchers prefer to spell out the graphical parameters of figures using caption descriptions such as ~~—e.g.~~ "Samples from Area A are shown with closed diamonds (◆, solid regression line), samples from Area B are shown with open diamonds (◇, dashed regression line)~~, etc.~~". ~~—w~~We argue that it is considerably easier ~~easiest~~ for readers to discern the aesthetic assignments of graphs by looking at legends rather than reading long descriptions in captions. Vaiglova et al. (2020b, Fig.2, in SM~~M~~ 1.E) provide an example of how nested legends can be used to describe the aesthetic assignments of several variables (species, chronological period, site, material measured) for separate panels, and for all panels combined.

4. Formal analysis: do not focus on statistical significances

For decades, statisticians and researchers in many fields of science ~~and statisticians alike~~ have decried the rote application of ~~the~~ null hypothesis ~~statistical~~ testing ~~(NHST) paradigm~~, specifically the aspect of it that compares a computed *p*-value to an arbitrary threshold (often 0.05 or 0.1) and declares the result "statistically significant" or not. The *p*-value itself is prone to many misconceptions (see for example Goodman 2008; Greenland et al. 2013), while statistical significance is often confused with practical significance~~,~~ (i.e., those results that are of scientific importance). This way of thinking has become entrenched to the point that publication can hinge on whether or not results achieve statistical significance, in turn opening the door to various questionable research practices such as *p*-hacking (Simmons et al. 2011; Head et al. 2015).

In 2016, the American Statistical Association (ASA) responded to these worrying and widespread misuses of statistical ~~techniques~~ testing by issuing a 'Statement on *p*-values' (Wasserstein and Lazar, 2016). The statement provides important clarifications on what *p*-values do and do n~~'~~ot measure, what they do and do n~~'~~ot represent, and what they should and should not be used for. The statement includes six principles; commentaries from the drafters of the statement (not all of whom agreed with the final version) are available as online supplements. Subsequently, in 2019, *The American Statistician* published a Special Issue on moving to a world beyond '$p<0.05$' with articles espousing a range of views on the controversies now swirling around this most common of statistical tools. The lead editorial to the Special Issue (Wasserstein et al. 2019) called for the abandonment of strict thresholding of statistical quantities (such as the *p*-value) as well as of the use of the term "statistically significant~~ee~~."

Drawing on these debates, this section aims to provide some of the most pressing clarifications on the uses and misuses of statistical significance testing in isotope archaeology. It outlines advice on using the most common statistical test applied to isotopic data: comparing meaningful differences between groups (e.g., human groups distinguished by age, sex, social status, type of burial, or between different plant or animal species). While we do not wish to advocate for the abolishment of ~~abolishing the use of the~~ *p*-values, we strongly reiterate the statistical community's advice against 1) using the term ~~"~~'statistically significant difference~~"~~' (and related practices such as using ~~related practices, such as using~~ boldface or asterisks in a table to ~~communicate~~ highlight ~~which differences are deemed "statistically significant"~~specific groups) and 2) employing ~~"~~'significance

thresholds[22] (such as 0.05 and 0.1), which can inaccurately distinguish between *p*-values of 0.051 and 0.049 (values that are, in fact, not meaningfully different).

4.1 How do we test for meaningful differences between groups?

In a study investigating ancient agricultural management at the Neolithic site of Kouphovouno, southern Greece, Vaiglova et al. (2020a, p. 43) assessed differences in $\delta^{13}C$ values between four main crop species (free-threshing wheat, hulled barley, einkorn and pea). The paper and reported:

The data for the four main species (where n > 1) are normally distributed (Shapiro-Wilk test, W = 0.95, p = 0.09) and homogenous (Levene's test, F(3,38) = 1.12, p = 0.35), so an ANOVA test was used to assess statistically significant differences between the mean $\delta^{13}C$ values of the four groups. The results indicate that there are significant differences (F(3,38) = 25.26, p < 0.01) and a post hoc Bonferroni test reveals that the differences are between free-threshing wheat and hulled barley (p < 0.01), free-threshing wheat and pea (p < 0.01) and einkorn and pea (p = 0.004).

While this example does conform to many current conventions on statistical reporting, it is incomplete in important ways. For example, the actual means and their standard errors are missing. Instead, the authors provide only test statistics and (sometimes thresholded) *p*-values. This leaves the reader unable to judge if the differences among the groups are of scientific interest and deserving of further exploration; or what magnitude of difference would ~~what differences would~~ even be meaningful in this context. Tests for normality and homogeneity (especially the former) are popular in many fields. Strictly speaking, they may not be necessary, since the *t*-test and analysis of variance are relatively robust to minor departures from assumptions under mild conditions (Miller 1986). Furthermore, even if preplanned as part of the analysis, these tests would still need to be included in any corrections for multiple testing (to preserve overall Type I error control), thereby detracting from the comparisons of main interest. And, of course, each decision made on the basis of a hypothesis test is an opportunity for additional errors.

Rather than focusing on the test statistics and their *p*-values, ~~—~~along with accompanying statements about statistical significance, ~~—~~a more appropriate approach would be to provide ~~1)~~ plots showing the data distribution in each group ~~for easy comparison and assessment of normality,~~ along with~~nd 2)~~ the point estimates (i.e. mean differences) and their associated confidence intervals

Formatted: Font: Italic

Formatted: Indent: Left: 0"

Formatted: Font: Italic

(CI) ~~for evaluation of meaningful differences between groups~~. ~~However, i~~Instead of focusing on whether or not these intervals include zero, the researchers should define and focus on the range of "compatible" values by stating, for example, what differences would be meaningful to interpret. In this way, a richer engagement with the data can be presented, replacing the dry and arbitrary assessments of statistical significance, or lack thereof.

Following this advice, a more accurate summary of the analysis of the

~~A more appropriate and accurate analysis of the~~ differences in $\delta^{13}C$ values between the crops at Kouphovouno ~~would be~~is as follows:

Figure 1 shows the distribution of the $\delta^{13}C$ values of the four main crops in the assemblage, where mean $\delta^{13}C$ values are –24.3 ± 0.6 ‰ (n=15) for barley; –22.7 ± 0.5 ‰ (n=17) for free-threshing wheat; –23.6 ± 0.8 ‰ (n=6) for einkorn; –25.0 ± 0.9 ‰ (n=7) for pea. Small sample sizes make it difficult to assess normality, however, the crops exhibit somewhat similar levels of variability. A one-way analysis of variance was conducted to compare the $\delta^{13}C$ values of the four crops; here, F(3,38)=25.26, with a p-value of 0.002, indicates that there are differences in the mean $\delta^{13}C$ values between the crop species.

To explore these differences further, ~~we computed~~ Tukey's HSD intervals were computed to control for post-hoc multiple comparisons. Results are shown in Figure 2 and Table 2. Pea appears to have lower values than free-threshing wheat, einkorn, and hulled barley. However, inspection of the confidence intervals in the comparison with hulled barley indicates that slightly larger values are also compatible with the collected data and the small number of observations in einkorn make it hard to draw conclusions. Hulled barley has lower values than free-threshing wheat and einkorn, though again the confidence intervals in the latter indicates that slightly larger values are also compatible with the collected data. Finally, free-threshing wheat has higher values than einkorn, although with the small number of observations of these groups, it is hard to be definitive.

Focusing on isotopically meaningful differences—which we define as > 1.0 ‰ in this context—and keeping in mind the small sample sizes, the distinctions between the two types of wheat (free-threshing wheat and einkorn) and peas can be explained by peas' higher demand for water during the ~~grain~~seed-filling period. As wheat and barley grown under the same watering conditions exhibit physiologically-determined differences (with barley having $\delta^{13}C$ values c. 1–2 ‰ lower),

Formatted: Font: Italic
Formatted: Indent: Left: 0"
Formatted: Font: Italic
Formatted: Font: Italic
Formatted: Font: Italic
Formatted: Font: Italic
Formatted: Font: Italic
Formatted: Font: Italic
Formatted: Font: Italic
Formatted: Font: Italic
Formatted: Font: Italic

the measured difference of –1.6 ‰ (CI: –2.3, –0.8) is compatible with the two crops growing under the same watering regime.

4.2 Further comments on reporting meaningful differences: decimal places and effect sizes

When reporting *p*-values and ~~, as well as~~ other values such as means or test statistics, the question often arises as to how ~~"~~precisely~~"~~ those results should be reported. For *p*-values, two or three digits after the decimal are usually sufficient, hence 0.049 is preferred to 0.049326. Little important knowledge is gained by the extra trailing digits. On the contrary, unnecessary digits may impart a false sense of precision. Likewise, there is very little to be learned by reporting a *p*-value of 10^{-16}, or other similarly small numbers, which are in essence artifacts of machine precision. For averages and test statistics, a similar logic applies, and we should refrain from implying more precision than is supported by the data themselves. Ehrenberg (1981) advocates rounding to "two effective digits" to balance the lightening in cognitive load and information loss. In addition, meaningful isotopic differences may not necessarily lie at the highest level of measurement precision, and sample homogeneity needs to be taken into account when deciding how many decimal places to report and interpret (see more in Section 5.3).

It is good habit to report effect sizes or other meaningful summaries of the data, with their accompanying degrees of freedom, in addition to test statistics ~~(with their accompanying degrees of freedom) and *p* values~~. These should have measures of uncertainty attached to them (e.g., ~~,~~ standard errors or standard deviations as appropriate). If the data have been transformed for analytical~~sis~~ purposes, it can ~~nonetheless~~ be helpful to bring the results back to the original scale for interpretation. As in the rewritten example above, it is advisable to discuss the effect sizes and the uncertainty in the context of the original data and the scientific questions, rather than simply stating differences ~~(e.g.)~~ to be ~~"~~'statistically significant'.~~"~~

5. Writing scientifically rigorous publications

5.1 Avoiding common misuses of terminology

To ensure that isotope research is easily ~~discernible~~ comprehensible to practitioners working across varying subfields of isotope geochemistry, it is important that the data be disseminated using widely accepted conventions and standardized definitions of terms. Coplen (2011) presents

the guidelines for reporting isotope ratio measurements established by the Commission on Isotopic Abundances and Atomic Weights of the International Union of Pure and Applied Chemistry and Roberts et al. (2017a) provides examples of the most common misuses of the terminology in archaeological research. Here we summarize some of the most ~~pressing points~~misused conventions:

1. Isotope 'delta' notation (e.g., $\delta^{13}C$, $\delta^{15}N$, $\delta^{18}O$) is defined as the *relative difference of isotope ratios* between samples and international standards. It is in effect a 'ratio of a ratio' and should by convention be referred to as a 'value' (and not a 'ratio', 'signature', 'composition' or 'fingerprint'). Notation that expresses the relative abundance of specific isotopes (e.g., $^{87}Sr/^{86}Sr$) should be referred to as a 'ratio~~s~~' (and not a 'value', ~~or~~ 'signature' or 'fingerprint'). ~~It is correct to discuss the 'stable carbon isotope composition of a sample', but not its 'δ^{13}C composition'.~~
2. The words 'enriched'/'higher' ~~'depleted'~~ and '~~higher'/'~~lower'/'depleted' ~~have distinct uses~~are not interchangeable. Stable isotope *values* (e.g., $\delta^{13}C$, $\delta^{15}N$, $\delta^{18}O$) and *ratios* (e.g., $^{87}Sr/^{86}Sr$) can be higher or lower, but they cannot be enriched or depleted. Enrichment and depletion describe the abundance of specific isotopes themselves (e.g., ^{13}C, ^{15}N, ^{18}O) and need to specify which samples are being compared to one another, e.g., ~~When describing a stable isotope value as 'enriched' or 'depleted', the isotope in question (e.g. ^{13}C, ^{15}N, ^{18}O) must be specified and the value compared to another value in which the abundance of that isotope differs. Thus~~ "Sample A is enriched in ^{15}N compared to Sample B". ~~It would be incorrect to describe a sample or a 'delta' value as 'enriched' without specifying the isotope and what the value is being compared to. 'Higher' and 'lower' are used to refer to the isotopic values themselves, as in "The δ^{15}N value of sample A is higher compared to the δ^{15}N value of sample B".~~
3. ‰ is permil (parts per thousand), and not equivalent to ppm (parts per million).

~~—~~The terms

~~— Lesser-adhered-to conventions that are often dismissed as too pedantic include:~~

~~1.~~4.~~Distinguishing between the terms~~ 'isotope' and 'isotopic' have distinct uses.~~, wherein~~ '~~i~~Isotope' is used when the word is modified (e.g., "The stable carbon isotope value") while~~and~~ 'isotopic' is used when it stands alone (e.g., "The isotopic composition" or "isotopic variability") (Sharp 2017).

Formatted: Numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0.25" + Indent at: 0.5"

2.5.Negative values should be denoted with ~~Using~~ an en (–) dash ~~instead of~~and not a hypthen (-) ~~to denote negative values~~ (Coplen 2011).

3.6.The ~~Italicizing the~~ 'delta' symbol in $\delta^{13}C$, $\delta^{15}N$, $\delta^{18}O$~~^{18}O~~ should be italicized as per the convention of denoting quantity symbols (Coplen 2011).

5.2 Transparency in measurement calibration and error estimation ~~methods and assessment of measurement uncertainty~~ are indispensable for establishing data reliability ~~quality~~

Two of the biggest inconsistencies in isotope archaeology ~~literature~~ are the frequency with which publications report how raw isotope values were converted to calibrated δ-(delta) values (even though the choice of calibration method can have a significant impact on the accuracy of the results; Pollard et al. 2011; Szpak et al. 2017) and the level of detail provided in assessing measurement reliability. Precision and accuracy are two distinct indices of measurement error, and both are important for establishing data reliability. Accuracy refers to the 'trueness' of the measurements (i.e. how close they are to the true values of the samples), while precision captures the reproducibility of the measurements (i.e. whether repeated measurements on one instrument produce the same results). Yet, most publications in isotope archaeology do not report these two indices, nor do they specify what calibration methods (e.g., one-point, two-point, multi-point calibration) were used to calculate the δ-values. In a review of 487 archaeological publications presenting original isotopic data, Szpak et al. (2017) found that 83% of the studies provided no information about how the δ-values were calibrated, 95% of the studies did not mention accuracy and only 4% of the studies distinguished between accuracy and precision. The authors concluded that "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" (Szpak et al. 2017, p. 612).

It is necessary to move away from the current reporting of measurement error that provides only relatively uninformative statements such as "'Measurement uncertainty for both $\delta^{13}C$ and $\delta^{15}N$ was < 0.1 ‰"'. To accurately calculate measurement precision and accuracy, both *calibration standards* and *check standards* should be used consistently in every analytical session and publications should report measurement uncertainty for particular ~~analytical sessions~~runs, rather than 'rolling averages' for a given laboratory (Szpak et al. 2017). This becomes particularly

important as archaeological isotope research moves towards larger meta-analyses, where distinctly different measurement errors might render smaller datasets incomparable. Based on recommendations from the Joint Committee for Guides in Metrology (2008) and a series of other studies (p.614), Szpak et al. (2017, p. 614) present a series of best practices for calibrating raw isotope values and calculating measurement uncertainty. They provide a user-friendly Microsoft Excel tool that can calculate precision, accuracy, and a pooled standard deviation of the two for particular analytical sessions, and offer an example paragraph and Online Supplement File showing how to fully report measurement error for a given project.

5.3 The number of decimal places reported should be consistent with both measurement uncertainty and sample homogeneity

Stable isotope values of unknown samples can only be as precise as the reference materials analyzed alongside them. Thus, if the measurement precision is reported as $\delta^{13}C = 0.1$ ‰, sample $\delta^{13}C$ values should only be reported to one decimal place (Bond and Hobson 2012). However, many archaeological materials do not present the same level of internal homogeneity as the reference materials. For example, in a study of inter- and intra-bone variability (assessed through measurement of multiple skeletal elements from animals representing a wide variety terrestrial, marine, and freshwater species), the mean intra-bone variability was 0.63 ± 0.06 ‰~~, while~~ and the mean inter-individual isotopic variation was 1.45 ± 1.15 ‰ (Hyland et al. 2021). Thus, even though precise $\delta^{13}C$ and $\delta^{15}N$ measurements can be obtained to two decimal places, the second decimal place may not provide any meaningful information. Authors should thus make sure that the reported values are consistent across a publication and that the reported error does not give a false impression about the interpretability of the data. ~~, and authors should be consistent in publications with reporting precision that is in line with their research questions.~~ In addition, elemental compositions (%C, %N) should be reported to one decimal place, while C:N ratios should be reported to two decimal places (see discussion in Guiry and Szpak 2021).

Strontium isotope measurements can be precise up to 5 or 6 decimal places (depending on the instrumentation) but in a study investigating intra-individual variation in $^{87}Sr/^{86}Sr$ ratios in human tooth enamel, Knudson et al. (2016) found that differences between individual measurements ranged from below detection level to 0.00015. They thus concluded that meaningful archaeological

interpretations about palaeomobility using $^{87}Sr/^{86}Sr$ ratios can be made at the third or fourth decimal place.

5.4 Writing with clarity and brevity

"*A scientific article that presents all of the data and all of the scientific discourse that the author intended to present is not necessarily a successful article. It only becomes one when most of the people who read it can perceive accurately and quickly what the author really meant.*" (Lindsay 2011, p. 64). Reading papers that are not written clearly and precisely is laborious and time consuming, and when pressed for time, we often end up not being able to read these papers in detail. This ~~may~~can result in papers that present important and interesting research not receiving the attention and credit they deserve.

Writing clearly and precisely is not necessarily a question of language proficiency. In his book *Scientific Writing = Thinking in Words*, David Lindsay (2011, p. 12) writes:

"*Science has a language of its own that has nothing to do with the scientist's native tongue. It is the language of logic in which reasoned arguments are developed from well-presented evidence and lead to sound and consistent conclusions. That language is the same regardless of the origin and preferred tongue of the person who writes it and good scientific writing depends primarily on expressing the science precisely and clearly. Subsequent editing by a native speaker to tidy up English expressions and comply with modern vernacular is relatively easy and the article will be a good one.*"

Numerous authors have written in detail on how to produce high-quality scientific writing (notably ~~(~~Alley 1996; O'Connor 2002; Gustavii 2016). Here we summarize advice from Lindsay (2011) on ~~provides concrete advice and examples on~~ how to build a logical structure for a scientific paper, how to achieve clarity and flow by presenting the relevant information in the right order, and how to make simple but powerful stylistic adjustments to enable the reader to grasp the meaning quickly and accurately. It is beyond the scope of this paper to summarize all of the key advice~~, but i~~In a plea for scientific practitioners—both those who are comfortable with writing, and those who find

Formatted: Font: Italic

Formatted: Indent: Left: 0"

Formatted: Font: Italic

writing challenging—to consult this and other resources~~book~~, we point out three main ~~messages that the author outlines:~~pieces of advice.

(a) The Introduction is the "powerhouse of your article" (p.20). It orientates the reader within the wider discussion to which the authors wish to contribute new knowledge. It does not just provide the theoretical or methodological background to the study.~~,~~ ~~but~~ It provides the foundation of the logical structure of the overarching argument and sets up the flow with which the Results and Discussion can be delivered effectively. To achieve these goals, writing the Introduction entails following two simple principles (p.21):

1. The hypothesis is key to the Introduction.
2. By justifying the hypothesis logically and scientifically, the writer provides just about everything necessary for readers to understand what the paper is about and why the author wrote it.

Although many writers prefer to write the Introduction last, given its central role in the logical structure of the argument, Lindsay (2011) convincingly demonstrates why it can be ~~can be~~ beneficial to start writing it first.

(b) The Introduction, Results, and Discussion work together. For the Discussion to be meaningful and convincing, the reader must be prepared (from the Introduction) to understand how the evidence gathered in this study (Results) logically feeds into the argument. To achieve their impact, results should be presented in the order of importance. Thus, rather than saving the most important for the end to create suspense, the most important finding should ~~thus~~ be presented first. To achieve this, Lindsay (2011) suggests that scientists sort their results into four categories (p.31):

1. Results that are clear and relevant to what you want to say about the hypothesis.
2. Results that allow you to say something relevant about the hypothesis but that are less convincing than the results in Category 1.
3. Results that are interesting, substantial, and worth presenting, but they don't have anything to do with the hypothesis.
4. Results that are not convincing and don't have anything to do with the hypothesis.

Once sorted, the results can be easily transformed into scaffolding for the logical argument. Category 4 results should be omitted, while Category 1–3 should be presented in exactly this order.

(c) Following the mantra that scientific writing should 'inform, not impress' will automatically result in avoiding some of the stumbling blocks that frequently characterize scientific writing" (p.5~~5~~6). This will solve two problems. Firstly, it will prevent the Discussion from becoming too long as a result of inclusion of unnecessary references and deliberations that do not lead to the conclusions being made (which, according to editors of scientific journals, are some of the most common problems of scientific manuscripts). Secondly, it will free the language from problems of clarity and brevity. Lindsay (2011) explains and provides examples for how to avoid seven~~7~~ main types of writing stumbling blocks: clusters of nouns; complex adjectival phrases; sentences beginning with subordinate clauses; using nouns instead of the verbs from which they are derived; use of imprecise words; use of acronyms; unfamiliar abbreviations, and symbols; footnotes, asides in parentheses, and other distractions. Examples of these stumbling blocks from isotope archaeology are discussed in Supplementary Material 3. In addition, the author lays out five~~5~~ steps which can be used to check the clarity of every paragraph: Is it a paragraph?; Do the sentences flow?; Are there stumbling blocks?; Can it be shortened without losing the meaning?; Does it say what you want it to say?. Due to space constraints, we are unable to provide examples for all these points of advice, but reiterate our conviction that both seasoned and up-and-coming writers will benefit from engaging with Lindsay's short book.

6. Conclusions

This paper ~~discussed~~ has presented suggestions and opportunities for improvement in three areas of project design and implementation: ~~sample selection~~selecting samples, ~~data processing~~processing data and ~~dissemination of results~~disseminating results. Some of the suggestions are relatively easy to implement:

- Consider carrying out power analyses to determine optimal sample sizes
- Invest time into exploring the distribution of data prior to formal statistical analysis
- Consider issues like overcrowding and colorblind-friendly palettes when preparing graphical figures
- Do not use the term ~~"~~'statistically significant difference~~"~~' (or any paraphrased equivalents)

- Do not hinge interpretation of group differences on *p*-value thresholds
- ~~Consider carrying out power analyses to determine optimal numbers of samples~~
- ~~Invest time into exploring the distribution of data prior to formal statistical analysis~~
- Consider measurement uncertainty and sample type when deciding how many decimal places to report; and report this number consistently throughout a publication
- ~~Consider issues like overcrowding and colorblind-friendly palettes when preparing figures~~
- Follow the established conventions and terminology for reporting stable isotope values ($\delta^{13}C$, $\delta^{15}N$, $\delta^{18}O$) and isotopic ratios ($^{87}Sr/^{86}Sr$)
- Take steps to justify sample selection, preservation and integrity (e.g.~~,~~ when reporting analyses of plants, include a supplementary document with photographs showing the surface morphology and internal structure of all analyzed seeds; when reporting the analyses of shells, bones and teeth, provide a table with results of all chemical analyses undertaken during sample screening)
- Consider how $\delta^{13}C$ values are obtained during AMS radiocarbon dating when deciding whether to use the values for answering palaeodietary/palaeoclimatic questions
- ~~Consider how $\delta^{13}C$ values are obtained during AMS radiocarbon dating when deciding whether to use the values for answering palaeodietary/palaeoclimatic questions~~

Some will require an initial investment in training, which will pay off in the future:

- Assess measurement error by calculating both precision and accuracy (using measured values of both calibration and check standards inserted into each analytical run) and provide all relevant details in supplementary materials to enable the reader to replicate the calculations independently
- Undertake targeted analyses of shell carbonate structure to determine their integrity prior to analysis
- Assess both the surface morphology and internal structure (shown in cross-sections) of charred seeds to determine their likelihood of preserving usable isotopic values

Some will require an ongoing investment in time and effort:

- Consult the evolving literature on the latest advice for quality control assessment
- Consult with professional statisticians on the design of experiments and collection of data
- Before analyzing crop species that have not been analyzed before, carry out charring experiments to determine the crops' preservation potential

Although isotopic research in archaeology ~~necessitates~~ requires a group effort and not everyone involved needs to understand the intricacies of the method, time investment from all parties involved can make a difference on both the final numbers and their interpretations. When researchers who cannot run analyses on their own are given full and transparent descriptions of the laboratory protocols and workflows from the collaborating laboratories and technicians, they will be better informed about where possible sources of error can occur. When archaeologists who have not been trained in isotopic geochemistry, but who are interested in collaborating on isotopic analyses, try their hand at data calibration and other data handling, they will be better positioned to communicate ~~how reliable the results are~~the reliability of the results. When students make time to practice (or at least engage in discussions ~~about~~surrounding) conventions and best practices such as those suggested in this paper, they will become effective in communicating the results both verbally and in writing.

Close collaboration with professional statisticians is crucial for avoiding ~~blunders~~ mistakes that have been, and are continuing to be made, by the wider scientific community. In addition, we encourage isotope practitioners to write rebuttal papers more frequently, to participate in sharing data on Open Access repositories (e.g., IsoArch, IsoMemo), and to interact with papers during the revision stage (by reading and commenting on 'pre-prints' online). Lastly, many people who speak to, collaborate with, or receive feedback from Julia Lee-Thorp have heard her say "Why would you want to do that?" in a variety of contexts. To ~~help keep her healthy dose of skepticism alive~~follow in the footsteps of her healthy dose of skepticism, we urge researchers to keep this question in mind during all stages of research. At no point should there be a time when ~~that when working on isotopic analyses, there is never a time when w~~they've cannot answer ~~this question~~it, or find out where the answer would be.

Author contributions

Conceptualization: PV, NL; Data curation: N/A; Formal analysis: PV, NL; Funding acquisition: N/A; Investigation: N/A; Methodology: N/A; Project administration: N/A; Resources: N/A; Software: N/A; Supervision: N/A; Validation: N/A; Visualization: PV, NL; Roles/Writing - original draft: PV, NL, ES, EL, CM~~, MB, RS~~; Writing - review & editing: PV, NL, ES, EL, CM~~, MB, RS~~.

Data availability

No original data was used in this paper. Stable isotope data that was re-examined in Section 5 was drawn from Vaiglova et al. (2020) and can be found at: https://link.springer.com/article/10.1007/s12520-019-00960-y#additional-information.

Declaration of competing interest

The authors declare no competing interests.

Acknowledgements

We are thankful to Richard Staff, David Chivall, Peter Ditchfield, Patrick Roberts, Maura Pellegrini, Melanie Beasley, Corrie Hyland, Rick Schulting, and Melissa Ritchey for their helpful comments and suggestions.

References

Aldeias V, Gur-Arieh S, Maria R, et al. (2019) Shell we cook it? An experimental approach to the microarchaeological record of shellfish roasting. Archaeol Anthropol Sci 11:389–407.
Alley M (1996) The craft of scientific writing. Springer, New York.
Alves EQ, Macario KD, Urrutia FP, et al. (2019) Accounting for the marine reservoir effect in radiocarbon calibration. Quat Sci Rev 209:129–138.
Ambrose SH (1990) Preparation and characterization of bone and tooth collagen for isotopic analysis. J Archaeol Sci 17:431–451.
Ambrose SH, Norr L (1993) Experimental Evidence for the Relationship of the Carbon Isotope Ratios of Whole Diet and Dietary Protein to Those of Bone Collagen and Carbonate. In: Lambert JB, Grupe G (eds) Prehistoric Human Bone: Archaeology at the Molecular Level. Springer, Berlin, pp 1–37.
Andrus CFT (2011) Shell midden sclerochronology. Quat Sci Rev 30:2892–2905.
Atalay S (2010) 'We don't talk about Çatalhöyük, we live it': sustainable archaeological practice through community-based participatory research. World Archaeol 42:418–429.
Aubert M, Williams IS, Boljkovac K, et al. (2012) In situ oxygen isotope micro-analysis of faunal material and human teeth using a SHRIMP II: a new tool for palaeo-ecology and archaeology. J Archaeol Sci 39:3184–3194.
Austin C, Smith TM, Farahani RMZ, et al. (2016) Uncovering system-specific stress signatures

in primate teeth with multimodal imaging. Sci Rep 6:18802.
Balasse M (2003) Potential biases in sampling design and interpretation of intra-tooth isotope analysis. Int J Osteoarchaeol 13:3–10.
Balasse M, Bălăşescu A, Janzen A, et al. (2013) Early herding at Măgura-Boldul lui Moş Ivănuş (early sixth millennium BC, Romania): environments and seasonality from stable isotope analysis. Eur J Archaeol 16:221–246.
Balasse M, Bocherens H, Mariotti A, Ambrose SH (2001) Detection of Dietary Changes by Intra-tooth Carbon and Nitrogen Isotopic Analysis: An Experimental Study of Dentine Collagen of Cattle (Bos taurus). J Archaeol Sci 28:235–245.
Balasse M, Boury L, Ughetto-Monfrin J, Tresset A (2012) Stable isotope insights ($\delta^{18}O$, $\delta^{13}C$) into cattle and sheep husbandry at Bercy (Paris, France, 4th millennium BC): birth seasonality and winter leaf foddering. Environ Archaeol 17:29–44.
Balasse M, Gillis R, Živaljević I, et al. (2021) Seasonal calving in European Prehistoric cattle and its impacts on milk availability and cheese-making. Sci Rep 11:1–11.
Beasley MM, Bartelink EJ, Taylor L, Miller RM (2014) Comparison of transmission FTIR, ATR, and DRIFT spectra: implications for assessment of bone bioapatite diagenesis. J Archaeol Sci 46:16–22.
Birch J (2012) Worldwide prevalence of red-green color deficiency. JOSA A 29:313–320.
Bogaard A, Heaton THE, Poulton P, Merbach I (2007) The impact of manuring on nitrogen isotope ratios in cereals: archaeological implications for reconstruction of diet and crop management practices. J Archaeol Sci 34:335–343.
Bonafini M, Pellegrini M, Ditchfield P, Pollard AM (2013) Investigation of the 'canopy effect' in the isotope ecology of temperate woodlands. J Archaeol Sci 40:3926–3935.
Bond AL, Hobson KA (2012) Reporting Stable-Isotope Ratios in Ecology: Recommended Terminology, Guidelines and Best Practices. Waterbirds 35:324–331.
Brand U (1989) Aragonite-calcite transformation based on Pennsylvanian molluscs. Geol Soc of Am Bull 101:377–390.
Cavazzuti C, Skeates R, Millard AR, et al. (2019) Flows of people in villages and large centres in Bronze Age Italy through strontium and oxygen isotopes. Plos One 14:e0209693.
Cernusak LA, Ubierna N, Winter K, et al. (2013) Environmental and physiological determinants of carbon isotope discrimination in terrestrial plants. New Phytol 200:950–965.
Charles M, Forster E, Wallace M, Jones G (2015) "Nor ever lightning char thy grain": establishing archaeologically relevant charring conditions and their effect on glume wheat grain morphology. STAR Sci Technol Archaeol Res 1:1–6.
Chesson LA, Beasley MM, Bartelink EJ, et al. (2021) Using bone bioapatite yield for quality control in stable isotope analysis applications. J Archaeol Sci Rep 35:102749.
Cheung C, Szpak P (2021) Interpreting Past Human Diets Using Stable Isotope Mixing Models—Best Practices for Data Acquisition. J Archaeol Method Theory 1–24.
Clementz MT (2012) New insight from old bones: stable isotope analysis of fossil mammals. J Mammal 93:368–380.
Cleveland WS (1994) The Elements of Graphing Data. Hobart Press, New Jersey.
Codron D, Hull J, Brink JS, et al (2011) Effect of competition on niche dynamics of syntopic grazing ungulates: contrasting the predictions of habitat selection models using stable isotope analysis. Evol Ecol Res 13:217–235.
Codron J, Lee-Thorp JA, Sponheimer M, Codron D (2013) Plant stable isotope composition across habitat gradients in a semi-arid savanna: implications for environmental reconstruction. J

Quat Sci 28:301–310.
Collard M, Wasef S, Adams S, et al. (2019) Giving it a burl: towards the integration of genetics, isotope chemistry, and osteoarchaeology in Cape York, Tropical North Queensland, Australia. World Archaeol 51:602–619.
Copeland SR, Sponheimer M, le Roux PJ, et al. (2008) Strontium isotope ratios ($^{87}Sr/^{86}Sr$) of tooth enamel: a comparison of solution and laser ablation multicollector inductively coupled plasma mass spectrometry methods. Rapid Commun Mass Spectrom 22:3187–3194.
Coplen TB (2011) Guidelines and recommended terms for expression of stable-isotope-ratio and gas-ratio measurement results. Rapid Commun Mass Spectrom 25:2538–560.
Cucchi T, Dai L, Balasse M, et al. (2016) Social Complexification and Pig (*Sus scrofa*) Husbandry in Ancient China: A Combined Geometric Morphometric and Isotopic Approach. Plos One 11:e0158523.
DeNiro MJ (1985) Postmortem preservation and alteration of in vivo bone collagen isotope ratios in relation to palaeodietary reconstruction. Nature 317:806–809.
DeNiro MJ, Weiner S (1988) Chemical, enzymatic and spectroscopic characterization of "collagen" and other organic fractions from prehistoric bones. Geochim Cosmochim Acta 52:2197–2206.
Dobberstein RC, Collins MJ, Craig OE, et al. (2009) Archaeological collagen: Why worry about collagen diagenesis? Archaeol Anthropol Sci 1:31–42.
Ehrenberg ASC (1981) The Problem of Numeracy. Am Stat 35:67–71.
Ericson JE (1985) Strontium isotope characterization in the study of prehistoric human ecology. J Hum Evol 14:503–514.
Evans JA, Montgomery J, Wildman G, Boulton N (2010) Spatial variations in biosphere $^{87}Sr/^{86}Sr$ in Britain. J Geol Soc 167:1–4.
Fernandes R (2016) A Simple(R) Model to Predict the Source of Dietary Carbon in Individual Consumers. Archaeometry 58:500–512.
Flohr P, Jenkins E, Williams HRS, et al. (2019) What can crop stable isotopes ever do for us? An experimental perspective on using cereal carbon stable isotope values for reconstructing water availability in semi-arid and arid environments. Veg Hist Archaeobotany 28:497–512.
France CAM, Sugiyama N, Aguayo E (2020) Establishing a preservation index for bone, dentin, and enamel bioapatite mineral using ATR-FTIR. J Archaeol Sci Rep 33:102551.
Fraser RA, Bogaard A, Heaton T, et al. (2011) Manuring and stable nitrogen isotope ratios in cereals and pulses: towards a new archaeobotanical approach to the inference of land use and dietary practices. J Archaeol Sci 38:2790–2804.
Frémondeau D, Cucchi T, Casabianca F, et al. (2012) Seasonality of birth and diet of pigs from stable isotope analyses of tooth enamel ($\delta^{18}O$, $\delta^{13}C$): a modern reference data set from Corsica, France. J Archaeol Sci 39:2023–2035.
Fry B (2006) Stable Isotope Ecology. Springer, Baton Rouge.
Galimberti M, Loftus E, Sealy J (2017) Investigating $\delta^{18}O$ of *Turbo sarmaticus* (L. 1758) as an indicator of sea surface temperatures. Palaeogeogr Palaeoclimatol Palaeoecol 484:62–69.
Gannes LZ, O'Brien DM, del Rio CM (1997) Stable Isotopes in Animal Ecology: Assumptions, Caveats, and a Call for More Laboratory Experiments. Ecology 78:1271–1276.
Gillis RE, Bulatović J, Penezić K, et al. (2021) Of herds and societies—Seasonal aspects of Vinča culture herding and land use practices revealed using sequential stable isotope analysis of animal teeth. Plos One 16:e0258230.
Gillis RE, Eckelmann R, Filipović D, et al. (2020) Stable isotopic insights into crop cultivation,

animal husbandry, and land use at the Linearbandkeramik site of Vráble-Veľké Lehemby (Slovakia). Archaeol Anthropol Sci 12:256.
Goodfriend GA (1992) The use of land snail shells in paleoenvironmental reconstruction. Quat Sci Rev 11:665–685.
Goodman S (2008) A Dirty Dozen: Twelve P-Value Misconceptions. Semin Hematol 45:135–140.
Greenland S, Senn SJ, Rothman KJ, et al. (2013) Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 31:337–350.
Guiry EJ, Szpak P (2021) Improved quality control criteria for stable carbon and nitrogen isotope measurements of ancient bone collagen. J Archaeol Sci 132:105416.
Guiry EJ, Szpak P, Richards MP (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.
Gustavii B (2016) How to Write and Illustrate a Scientific Paper, 2nd edn. Cambridge University Press, Cambridge.
Hart JP, Feranec RS (2020) Using Maize δ^{15}N values to assess soil fertility in fifteenth- and sixteenth-century ad Iroquoian agricultural fields. Plos One 15:e0230952.
Head ML, Holman L, Lanfear R, et al. (2015) The Extent and Consequences of P-Hacking in Science. Plos Biol 13:e1002106.
Hillson S (2005) Teeth. Cambridge university press, Cambridge.
Holt E, Evans JA, Madgwick R (2021) Strontium (^{87}Sr/^{86}Sr) mapping: A critical review of methods and approaches. Earth-Sci Rev 216:103593.
Hoppe KA, Stover SM, Pascoe JR, Amundson R (2004) Tooth enamel biomineralization in extant horses: implications for isotopic microsampling. Palaeogeogr Palaeoclimatol Palaeoecol 206:355–365.
Hyland C, Scott MB, Routledge J, Szpak P (2021) Stable Carbon and Nitrogen Isotope Variability of Bone Collagen to Determine the Number of Isotopically Distinct Specimens. J Archaeol Method Theory.
Ivany LC (2012) Reconstructing Paleoseasonality from Accretionary Skeletal Carbonates—Challenges and Opportunities. Paleontol Soc Pap 18:133–166.
Janzen A, Balasse M, Ambrose SH (2020) Early pastoral mobility and seasonality in Kenya assessed through stable isotope analysis. J Archaeol Sci 117:1–17.
Johnson BJ, Fogel ML, Miller GH (1998) Stable isotopes in modern ostrich eggshell: a calibration for paleoenvironmental applications in semi-arid regions of southern Africa. Geochim Cosmochim Acta 62:2451–2461.
Joint Committee for Guides in Metrology (2008) Evaluation of Measurement Data – Guide to the Expression of Uncertainty in Measurement. International Organization for Standardization, Geneva.
Jones AM, O'Connell TC, Young ED, et al. (2001) Biogeochemical data from well preserved 200 ka collagen and skeletal remains. Earth Planet Sci Lett 193:143–149.
Jones PJ, O'Connell TC, Jones MK, et al (2021) Crop water status from plant stable carbon isotope values: A test case for monsoonal climates. The Holocene 31:993–1004.
Knipper C, Pichler SL, Rissanen H, et al. (2017) What is on the menu in a Celtic town? Iron Age diet reconstructed at Basel-Gasfabrik, Switzerland. Archaeol Anthropol Sci 9:1307–1326.
Knipper C, Reinhold S, Gresky J, et al. (2020) Diet and subsistence in Bronze Age pastoral communities from the southern Russian steppes and the North Caucasus. Plos one 15:e0239861.

Knudson KJ, Stanish C, Lozada Cerna MC, et al. (2016) Intra-individual variability and strontium isotope measurements: A methodological study using $^{87}Sr/^{86}Sr$ data from Pampa de los Gentiles, Chincha Valley, Peru. J Archaeol Sci Rep 5:590–597.
Koch PL, Tuross N, Fogel ML (1997) The Effects of Sample Treatment and Diagenesis on the Isotopic Integrity of Carbonate in Biogenic Hydroxylapatite. J Archaeol Sci 24:417–429.
Lee-Thorp J, Sponheimer M (2003) Three case studies used to reassess the reliability of fossil bone and enamel isotope signals for paleodietary studies. J Anthropol Archaeol 22:208–216.
Lee-Thorp J, Van der Merwe NJ (1987) Carbon isotope analysis of fossil bone apatite. South Afr J Sci 83:712–715.
Lee-Thorp JA (2008) On isotopes and old bones. Archaemetry 50:925–950.
Leng MJ, Lewis JP (2014) Oxygen isotopes in Molluscan shell: Applications in environmental archaeology. Environ Archaeol 21:295–306.
Levin NE, Brown FH, Behrensmeyer AK, et al. (2011) Paleosol carbonates from the Omo Group: Isotopic records of local and regional environmental change in East Africa. Palaeogeogr Palaeoclimatol Palaeoecol 307:75–89.
Lindsay D (2011) Scientific Writing = Thinking in Words. Csiro Publishing, Collingwood.
Loftus E, Rogers K, Lee-Thorp J (2015) A simple method to establish calcite:aragonite ratios in archaeological mollusc shells. J Quat Sci 30:731–735.
Loftus E, Rogers K, Lee-Thorp J (2015) A simple method to establish calcite:aragonite ratios in archaeological mollusc shells. J Quat Sci 30:731–735.
Lowenstam HA (1954) Factors Affecting the Aragonite: Calcite Ratios in Carbonate-Secreting Marine Organisms. J Geol 62:284–322.
Macfadden BJ, Cerling TE (1996) Mammalian herbivore communities, ancient feeding ecology, and carbon isotopes: A 10 million-year sequence from the Neogene of Florida. J Vertebr Paleontol 16:103–115.
Makarewicz CA, Pederzani S (2017) Oxygen ($\delta^{18}O$) and carbon ($\delta^{13}C$) isotopic distinction in sequentially sampled tooth enamel of co-localized wild and domesticated caprines: Complications to establishing seasonality and mobility in herbivores. Palaeogeogr Palaeoclimatol Palaeoecol 485:1–15.
Marston JM (2021) Archaeological Approaches to Agricultural Economies. J Archaeol Res 29:327–385.
Masters PM (1987) Preferential preservation of noncollagenous protein during bone diagenesis: Implications for chronometric and stable isotopic measurements. Geochim Cosmochim Acta 51:3209–3214.
Milano S, Prendergast AL, Schöne BR (2016) Effects of cooking on mollusk shell structure and chemistry: Implications for archeology and paleoenvironmental reconstruction. J Archaeol Sci Rep 7:14–26.
Miller RG (1986) Beyond ANOVA: Basics of Applied Statistics. John Wiley and Sons, New York.
Müldner G, Richards MP (2007) Stable isotope evidence for 1500 years of human diet at the city of York, UK. Am J Phys Anthropol 133:682–697.
Murphy KR, Myors B, Wolach A (2014) Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests, Fourth Edition, 4th edn. Routledge, New York.
Nitsch E, Andreou S, Creuzieux A, et al. (2017) A bottom-up view of food surplus: using stable carbon and nitrogen isotope analysis to investigate agricultural strategies and diet at Bronze Age

Archontiko and Thessaloniki Toumba, northern Greece. World Archaeol 49:105–137.
Nitsch EK, Charles M, Bogaard A (2015) Calculating a statistically robust $\delta^{13}C$ and $\delta^{15}N$ offset for charred cereal and pulse seeds. Sci Technol Archaeol Res 1:1–8.
O'Connor M (2002) Writing Successfully in Science. Routledge, London.
Pálsdóttir AH, Bläuer A, Rannamäe E, et al. (2019) Not a limitless resource: ethics and guidelines for destructive sampling of archaeofaunal remains. R Soc Open Sci 6:191059.
Parnell AC, Phillips DL, Bearhop S, et al. (2013) Bayesian stable isotope mixing models. Environmetrics 24:387–399.
Pearson J, Grove M (2013) Counting sheep: sample size and statistical inference in stable isotope analysis and palaeodietary reconstruction. World Archaeol 45:373–387.
Pechenkina EA, Ambrose SH, Xiaolin M, Benfer RA (2005) Reconstructing northern Chinese Neolithic subsistence practices by isotopic analysis. J Archaeol Sci 32:1176–1189.
Pederzani S, Britton K (2019) Oxygen isotopes in bioarchaeology: Principles and applications, challenges and opportunities. Earth-Sci Rev 188:77–107.
Pollard AM, Pellegrini M, Lee-Thorp JA (2011) Technical note: Some observations on the conversion of dental enamel $\delta^{18}O_p$ values to $\delta^{18}O_w$ to determine human mobility. Am J Phys Anthropol 145:499–504.
Price M, Rowan YM, Kersel MM, Makarewicz CA (2020) Fodder, pasture, and the development of complex society in the Chalcolithic: isotopic perspectives on animal husbandry at Marj Rabba. Archaeol Anthropol Sci 12:95.
Pryor AJE, Insoll T, Evis L (2020) Laser ablation strontium isotope analysis of human remains from Harlaa and Sofi, eastern Ethiopia, and the implications for Islamisation and mobility. STAR Sci Technol Archaeol Res 6:113–136.
Reimer PJ, Brown TA, Reimer RW (2004) Discussion: Reporting and Calibration of Post-Bomb 14C Data. Radiocarbon 46:1299–1304.
Reyes-Gasga J, Martínez-Piñeiro EL, Rodríguez-Álvarez G, et al. (2013) XRD and FTIR crystallinity indices in sound human tooth enamel and synthetic hydroxyapatite. Mater Sci Eng C 33:4568–4574.
Roberts P, Fernandes R, Craig OE, et al. (2017a) Calling all archaeologists: guidelines for terminology, methodology, data handling, and reporting when undertaking and reviewing stable isotope applications in archaeology. Rapid Commun Mass Spectrom 32:361–372.
Roberts P, Perera N, Wedage O, et al. (2017b) Fruits of the forest: Human stable isotope ecology and rainforest adaptations in Late Pleistocene and Holocene (~36 to 3 ka) Sri Lanka. J Hum Evol 106:102–118.
Roberts P, Gillingwater TH, Lahr MM, et al (2018) Historical Tropical Forest Reliance amongst the Wanniyalaeto (Vedda) of Sri Lanka: an Isotopic Perspective. Hum Ecol 46:435–444.
Roberts P, Perera N, Wedage O, et al. (2015) Direct evidence for human reliance on rainforest resources in late Pleistocene Sri Lanka. Science 347:1246–1249.
Roffet-Salque M, Marciniak A, Valdes PJ, et al. (2018) Evidence for the impact of the 8.2-kyBP climate event on Near Eastern early farmers. Proc Natl Acad Sci 115:8705–8709.
Rokita E, Hermes C, Nolting H-F, Ryczek J (1993) Substitution of calcium by strontium within selected calcium phosphates. J Cryst Growth 130:543–552.
Royer A, Daux V, Fourel F, Lécuyer C (2017) Carbon, nitrogen and oxygen isotope fractionation during food cooking: Implications for the interpretation of the fossil human record. Am J Phys Anthropol 163:759–771.
Rull V, Vegas-Vilarrúbia T, Corella JP, Valero-Garcés B (2021) Bronze Age to Medieval

vegetation dynamics and landscape anthropization in the southern-central Pyrenees. Palaeogeogr Palaeoclimatol Palaeoecol 571:110392.
Schoeninger MJ, Moore KM, Murray ML, Kingston JD (1989) Detection of bone preservation in archaeological and fossil samples. Appl Geochem 4:281–292.
Schöne BR (2008) The curse of physiology—challenges and opportunities in the interpretation of geochemical data from mollusk shells. Geo-Mar Lett 28:269–285.
Sharp Z (2017) Principles of Stable Isotope Geochemistry. Pearson, online only edition.
Shemesh A (1990) Crystallinity and diagenesis of sedimentary apatites. Geochim Cosmochim Acta 54:2433–2438.
Simmons JP, Nelson LD, Simonsohn U (2011) False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. Psychol Sci 22:1359–1366.
Smith TM, Austin C, Green DR, et al (2018) Wintertime stress, nursing, and lead exposure in Neanderthal children. Sci Adv 4:1–9.
Smith CI, Nielsen-Marsh CM, Jans MME, Collins MJ (2007) Bone diagenesis in the European Holocene I: patterns and mechanisms. J Archaeol Sci 34:1485–1493.
Snoeck C, Lee-Thorp J, Schulting R, et al. (2015) Calcined bone provides a reliable substrate for strontium isotope ratios as shown by an enrichment experiment. Rapid Commun Mass Spectrom 29:107–114.
Snoeck C, Ryan S, Pouncett J, et al. (2020) Towards a biologically available strontium isotope baseline for Ireland. Sci Total Environ 712:136248.
Snoeck C, Schulting RJ, Lee-Thorp JA, et al. (2016) Impact of heating conditions on the carbon and oxygen isotope composition of calcined bone. J Archaeol Sci 65:32–43.
Sponheimer M, Passey BH, de Ruiter DJ, et al. (2006) Isotopic Evidence for Dietary Variability in the Early Hominin Paranthropus robustus. Science 314:980–982.
Squires K, Booth T, Roberts CA (2019) The Ethics of Sampling Human Skeletal Remains for Destructive Analyses. Ethical Approaches Hum Remains 265–297.
Stevens RE, Lister AM, Hedges REM (2006) Predicting diet, trophic level and palaeoecology from bone stable isotope analysis: a comparative study of five red deer populations. Oecologia 149:12–21.
Stuiver M, Polach HA (1977) Discussion Reporting of ^{14}C Data. Radiocarbon 19:355–363.
Styring A, Maier U, Stephan E, et al. (2016) Cultivation of choice: new insights into farming practices at Neolithic lakeshore sites. Antiquity 90:95–110.
Styring AK, Charles M, Fantone F, et al. (2017) Isotope evidence for agricultural extensification reveals how the world's first cities were fed. Nat Plants 3:17076.
Styring AK, Evans JA, Nitsch EK, et al. (2019) Revisiting the potential of carbonized grain to preserve biogenic $^{87}Sr/^{86}Sr$ signatures within the burial environment. Archaeometry 61:179–193.
Sullivan CH, Krueger HW (1981) Carbon isotope analysis of separate chemical phases in modern and fossil bone. Nature 292:333–335.
Syväranta J, Lensu A, Marjomäki TJ, et al. (2013) An Empirical Evaluation of the Utility of Convex Hull and Standard Ellipse Areas for Assessing Population Niche Widths from Stable Isotope Data. Plos One 8:e56094.
Szpak P, Chiou KL (2019) A comparison of nitrogen isotope compositions of charred and desiccated botanical remains from northern Peru. Veg Hist Archaeobotany 29:527–538.
Szpak P, Metcalfe JZ, Macdonald RA (2017) Best practices for calibrating and reporting stable isotope measurements in archaeology. J Archaeol Sci Rep 13:609–616.

Thomson JJ, Aston FW, Soddy F, et al. (1921) Discussion on Isotopes. Proc R Soc Lond Ser Contain Pap Math Phys Character 99:87–104.
Tieszen LL (1991) Natural variations in the carbon isotope values of plants: implications for archaeology, ecology, and paleoecology. J Archaeol Sci 18:227–248.
Vaiglova P, Gardeisen A, Buckley M, et al. (2020a) Further insight into Neolithic agricultural management at Kouphovouno, southern Greece: expanding the isotopic approach. Archaeol Anthropol Sci 12:43.
Vaiglova P, Hartman G, Marom N, et al. (2020b) Climate stability and societal decline on the margins of the Byzantine empire in the Negev Desert. Sci Rep 10:1512.
Vaiglova P, Halstead P, Pappa M, et al. (2018) Of cattle and feasts: Multi-isotope investigation of animal husbandry and communal feasting at Neolithic Makriyalos, northern Greece. Plos One 13:e0194474.
Vaiglova P, Reid REB, Lightfoot E, et al. (2021) Localized management of non-indigenous animal domesticates in Northwestern China during the Bronze Age. Sci Rep 11:15764.
Vaiglova P, Snoeck C, Nitsch E, et al. (2014) Impact of contamination and pre- treatment on stable carbon and nitrogen isotopic composition of charred plant remains. Rapid Commun Mass Spectrom 28:2497–2510.
van Klinken GJ (1999) Bone Collagen Quality Indicators for Palaeodietary and Radiocarbon Measurements. J Archaeol Sci 26:687–695.
Vogel JC, van der Merwe NJ (1977) Isotopic Evidence for Early Maize Cultivation in New York State. Am Antiq 42:238–242.
Wainer H (2007) Graphic Discovery: A Trout in the Milk and Other Visual Adventures. Princeton University Press, Princeton.
Wallace M, Jones G, Charles M, et al. (2013) Stable carbon isotope analysis as a direct means of inferring crop water status and water management practices. World Archaeol 45:388–409.
Wang X, Roberts P, Tang Z, et al. (2021) The Circulation of Ancient Animal Resources Across the Yellow River Basin: A Preliminary Bayesian Re-evaluation of Sr Isotope Data From the Early Neolithic to the Western Zhou Dynasty. Front Ecol Evol 9:16.
Wasserstein RL, Lazar NA (2016) The ASA Statement on p-values: Context, Process, and Purpose. Am Stat 70:129–133.
Wasserstein RL, Schirm AL, Lazar NA (2019) Moving to a World Beyond "p < 0.05." Am Stat 73:1–19.
Webb EC, White CD, Longstaffe FJ (2014) Investigating inherent differences in isotopic composition between human bone and enamel bioapatite: implications for reconstructing residential histories. J Archaeol Sci 50:97–107.
Weber K, Weber M, Menneken M, et al. (2021) Diagenetic stability of non-traditional stable isotope systems (Ca, Sr, Mg, Zn) in teeth – An in-vitro alteration experiment of biogenic apatite in isotopically enriched tracer solution. Chem Geol 572:120196.
Weiner S, Bar-Yosef O (1990) States of preservation of bones from prehistoric sites in the Near East: A survey. J Archaeol Sci 17:187–196.
Weinreb MM (1964) Tooth development in sheep. Am J Vet Res 25:891–908.
Weissgerber TL, Winham SJ, Heinzen EP, et al. (2019) Reveal, Don't Conceal. Circulation 140:1506–1518.
Wickham H (2016) ggplot2: elegant graphics for data analysis 2016, 2nd edn. Springer.
Wierzbowski H, Anczkiewicz R, Pawlak J, et al. (2017) Revised Middle–Upper Jurassic strontium isotope stratigraphy. Chem Geol 466:239–255.

Wilkinson L (2005) The Grammar of Graphics. In: Handbook of Computational Statistics: Concepts and Methods, 2nd edn. Springer, New York.
Wood R, Fleury ABC, Fallon S, et al. (2021) Do weak or strong acid remove carbonate contamination from ancient tooth enamel more effectively? The effect of acid pretreatment on radiocarbon and $\delta^{13}C$ analyses. Radiocarbon 63:935–952.
Wright LE, Schwarcz HP (1996) Infrared and Isotopic Evidence for Diagenesis of Bone Apatite at Dos Pilas, Guatemala: Palaeodietary Implications. J Archaeol Sci 23:933–944.
Yanes Y, Izeta AD, Cattáneo R, et al. (2014) Holocene (~4.5–1.7 cal. kyr BP) paleoenvironmental conditions in central Argentina inferred from entire-shell and intra-shell stable isotope composition of terrestrial gastropods. The Holocene 24:1193–1205.
Yanes Y, Yapp CJ, Ibáñez M, et al. (2011) Pleistocene–Holocene environmental change in the Canary Archipelago as inferred from the stable isotope composition of land snail shells. Quat Res 75:658–669.
Zazzo A, Balasse M, Patterson WP (2005) High-resolution $\delta^{13}C$ intratooth profiles in bovine enamel: Implications for mineralization pattern and isotopic attenuation. Geochim Cosmochim Acta 69:3631–3642.
Zazzo A, Lécuyer C, Sheppard SMF, et al. (2004) Diagenesis and the reconstruction of paleoenvironments: A method to restore original $\delta^{18}O$ values of carbonate and phosphate from fossil tooth enamel. Geochim Cosmochim Acta 68:2245–2258.
Zhang Q, Hou Y, Li X, et al (2021) Stable isotopes reveal intensive pig husbandry practices in the middle Yellow River region by the Yangshao period (7000–5000 BP). Plos One 16:e0257524.

Captions

Figure 1 Distribution of $\delta^{13}C$ values of the four main crops at Kouphovouno. Data was originally published in Vaiglova et al. (2020a) and the statistical analysis is being re-examined here.

Figure 2 Outcome of Tukey HSD comparison between the $\delta^{13}C$ values of the four main crops at Kouphovouno. Data was originally published in Vaiglova et al. (2020a) and the statistical analysis is being re-examined here.

Table 1 Summary of ~~uggested~~ quality control criteria for ~~stable isotope analysis of most commonly used~~ archaeological materials most commonly analyzed in isotopic archaeology. Method development on these criteria is an active area of research, and the reader is encouraged to check the latest literature. References present experimental studies on quality control, unless preceded by "e.g.", in which case they present an example of an application of these criteria. ~~The C:N atomic ratio accounts for the varying atomic weights of the two elements and is calculated as follows: ((%C/12)/(%N/14)).~~

Table 2 Results of Tukey HSD comparisons between $\delta^{13}C$ values of free-threshing wheat, hulled barley, einkorn and pea. Data was originally published in Vaiglova et al. (2020a) and the statistical analysis is being re-examined here.

Revised Manuscript with changes accepted
Click here to view linked References

Best practices for selecting samples, analyzing data, and publishing results in isotope archaeology

Petra Vaiglova[a,*], Nicole A. Lazar[b], Elizabeth A. Stroud[c], Emma Loftus[d], Cheryl A. Makarewicz[e]

[a] Department of Anthropology, Washington University in St Louis, 1 Brookings Drive, St. Louis, MO, 63130, USA
[b] Department of Statistics, Penn State University, 326 Thomas Building, University Park, PA, 16802, USA
[c] School of Archaeology, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, UK
[d] Institute for Pre- and Protohistoric Archaeology and Provincial Roman Archaeology, Ludwig Maximilian University Munich, Geschwister-Scholl-Platz 1, 80539 Munich, Germany
[e] Institute of Prehistoric and Protohistoric Archaeology, University of Kiel, Johanna-Mestorf-Straße 2-6, Kiel, Germany

* Corresponding author.
E-mail address: vaiglova.petra@gmail.com (P. Vaiglova).

Keywords

Isotope archaeology; Palaeodiet; Palaeoclimate; Ancient mobility; Quality control; P-values

Abstract

Isotopic analysis has become one of the most popular arenas of archaeological science, in part due to its versatility to uncover intriguing insights from a range of organic and inorganic archaeological materials. However, alongside an increase in popularity, the field has seen the rise of dissemination of publications that do not pass quality control, do not apply robust interpretative frameworks, or do not report data in ways that would make them amenable to critical evaluation or inclusion in large meta-analyses. This paper represents an effort to clarify some of the most pressing weaknesses and misconceptions in 'traditional' applications of isotopic techniques in archaeology: measurement of stable carbon, nitrogen, and oxygen isotope values of organic and inorganic materials (bulk bone collagen, bulk dentine, and seeds; bulk and incremental tooth enamel, and molluscan shells), and strontium isotope ratio analysis of tooth enamel and cremated bone. The discussion centers on three key aspects of research:

(1) Selecting samples, with advice on building comparative baselines (or more appropriately 'base*intervals*') and words of caution on interpreting stable carbon isotope values measured during AMS radiocarbon dating.
(2) Handling data, including tips on exploratory data analysis, graphical visualization, and statistical assessment of differences between groups; with particular reference to the Statement on *p*-values published by the American Statistical Association.
(3) Reporting data, with advice on using correct terminology and decimal points, calculating measurement precision and accuracy, and communicating results using effective scientific language.
The advice provided in this paper does not cover all aspects of project design and dissemination, but will hopefully provide clarification within these three key areas and inspire further discussion of effective and impactful applications of isotopic techniques in archaeology.

1. Introduction

Since J.J.Thomson et al. (1921) first discussed the concept of elemental isotopes 100 years ago, isotopic analyses have gained traction in a wide array of fields including geochemistry, climatology, hydrology, ecology, plant physiology, palaeobiology, bacteriology, and archaeology (Sharp 2017). The first isotopic analyses of archaeological remains in the 1970s aimed at identifying human consumption of non-native maize in North America (Vogel and van der Merwe 1977). Not long after, Ericson (1985) proposed that strontium isotope ratios could be used to trace the geographical origins of human remains. However, it was not until foundational strides were made towards understanding tissue diagenesis and preservation of *in vivo* isotopic composition of archaeological materials (Sullivan and Krueger 1981; Lee-Thorp and Van der Merwe 1987; Masters 1987; Schoeninger et al. 1989), alongside advances in understanding biochemical mechanisms such as routing of dietary isotopes (e.g., Ambrose and Norr 1993) and incorporation of strontium into mineral bioapatite (e.g., Rokita et al. 1993), that isotope archaeology was on track to becoming one of the most popular branches of archaeological science. Today, its wide-ranging applications encompass both 'traditional' and rather inexpensive analyses of bulk stable isotope values (e.g., bulk bone collagen, bulk and incremental tooth enamel), as well as more instrumentally challenging and costlier analyses (e.g., compound specific stable isotope analysis of bone collagen amino acids, high-resolution elemental mapping of tooth cross-sections).

The growth of isotope archaeology was in large part driven by a generation of women scientists including Julia Lee-Thorp, Judith Sealy, Margaret Schoeninger, Noreen Tuross, Marilyn Vogel, and M. Anne Katzenberg. Through their perseverance and adherence to high standards of scientific rigor, these scientists trained a long list of students and produced work that would come to transform the way we ask questions about the human past. Forty five years after the birth of isotope archaeology, the core principles that characterize these researchers' work—asking meaningful questions, identifying reliable samples, building robust interpretative frameworks, and producing scientifically rigorous publications—remain the cornerstones of impactful research in this field.

This paper forms part of a Special Issue dedicated to the career and mentorship of Julia Lee-Thorp and aims to synthesize useful advice on obtaining and working with isotopic data in archaeological research. Through collaboration with archaeologists and palaeoecologists, Julia Lee-Thorp made foundational strides towards understanding the outcome and the process of tissue diagenesis and applying isotopic analyses for understanding early hominin diets and past human–ecology dynamics. As it is not possible to cover in sufficient detail all aspects of project design in one paper, we have chosen to focus on three key areas: sample selection, data processing, and data reporting. The discussion focuses on the most common applications: stable carbon, nitrogen, and oxygen isotope values of organic and inorganic materials (bulk bone collagen, bulk dentine, and seeds; bulk and incremental tooth enamel, and molluscan shells), and strontium isotope ratio analysis of tooth enamel and cremated bone. The paper does not aim to expose errors by any particular researchers, nor does it try to generalize about trends in the amount of biochemical and geochemical training that practitioners of isotope archaeology receive. It is simply aimed at outlining where improvements can be made in the production of meaningful and interesting research.

In some parts, this paper reiterates advice presented in Roberts et al.'s (2017a) '*Calling all archaeologists: guidelines for terminology, methodology, data handling, and reporting when undertaking and reviewing stable isotope applications in archaeology'* and Szpak et al.'s (2017) '*Best practices for calibrating and reporting stable isotope measurements in archaeology*'. This

paper is meant to serve as a companion to these publications (and others including Fry 2006; Bond and Hobson 2012; Sharp 2017), while providing some more recent advice on:

- Assessing sample quality for collagen (Guiry and Szpak 2021), charred seeds (Szpak and Chiou 2019), bioapatite (France et al. 2020) and shell (Loftus et al. 2015)
- Inspecting both surface and internal structure of grains to select charred plant material likely to provide reliable isotopic measurements (Stroud et al. in prep).
- Identifying archaeological reference samples with high likelihood of approximating the unaltered local isotopic baseline (or base*interval*) values.
- Determining the suitability of stable carbon isotope values measured during AMS radiocarbon dating for palaeodietary and palaeoclimate research.
- Effectively using exploratory data analysis and graphical visualization techniques to guide statistical analysis and effective presentation of data.
- Reporting statistical analyses in line with the American Statistical Association's recent Statement on *p*-values (Wasserstein and Lazar 2016; Wasserstein et al. 2019).

For a review of basic principles and research questions in isotope archaeology, we refer the reader to Julia Lee-Thorp's paper *On isotopes and old bones* (2008).

2. Identifying reliable samples

A mass spectrometer will always produce a measurement, no matter the sample preservation or the technical expertise of the operator. For this reason, it is important that researchers invest time in selecting samples that are likely to have retained their *in vivo* (i.e. original) isotopic composition. In this section, we discuss considerations of sample size, sample type, sampling contexts, and preservation potential for the main materials analyzed in isotope archaeology: bones, teeth, seeds, and shells.

Equally important as selecting reliable samples is to confirm that the acquired measurements are reliable. This can be accomplished, with caveats, by assessing the degree to which the chemical composition of the archaeological samples resembles the composition of analogous modern material (Zazzo et al. 2004; Beasley et al. 2014; Loftus et al. 2015; Szpak and Chiou 2019; Guiry and Szpak 2021). Table 1 provides a summary of currently accepted or suggested quality control

(QC) benchmarks for collagen, bone bioapatite, tooth enamel bioapatite, charred plant remains, and molluscan shell. It is important to recognize, however, that these criteria are not perfect, and will continue to be refined through further development (see discussion in Cheung and Szpak 2021).

It is also important to note that sample selection requires establishing a compromise between adequate sample size and ethical responsibility to preserve irreplaceable archaeological materials (Pálsdóttir et al. 2019; Squires et al. 2019). Curators are professionally obliged to safeguard archaeological collections and deny access to large sample numbers that may offer little in the way of improved resolution. To justify destructive sampling, researchers should outline reasonable expectations surrounding the isotopic variability in the local ecology and support their desired numbers of samples with statistically meaningful estimates of optimal sample size (carried out using, for example, power analyses). In addition, intentions to sample human remains must not only comply with relevant laws and regulations, but demonstrate the researchers' commitment to responsibly handle sensitive material and involve local communities in the research, particularly when it relates to descendant groups or controversial political discussions (Atalay 2010; Roberts et al. 2018; Collard et al. 2019).

2.1 Bones

Sampling bones for palaeodietary studies needs to strike a balance between the desired level of detail and the minimum number of samples that will provide a 'best case scenario' for approximating the means and variability within a given population or group. The level of attainable detail is constrained by the research question and the isotope variability characterizing the relevant ecosystem. For example, a project that aims to assess the overall dietary composition and homogeneity of a given population and a project that aims to assess possible differences in dietary intake between distinct sub-groups (identified by, for example, sex, age, pathologies, or association with distinct burial types, household units or site clusters) (e.g., Müldner and Richards 2007) will delineate different optimal sample sizes.

Power analysis can be used to estimate the minimum number of samples required to answer a research question (Murphy et al. 2014). It can be used to perform sample size calculations for a

range of reasonable standard deviations in order to determine how much the outcomes change as a result. This information can be used to establish the optimal sample numbers that will be amenable for answering the research question while avoiding redundancy (Pechenkina et al. 2005; Stevens et al. 2006). The indices required to carry out these calculations (likely effect sizes, group differences, and measures of variability such as standard errors) can be obtained from the literature or through pilot studies. They do not have to be precise, since the goal is to obtain a general idea of the sample sizes needed to detect, for example, differences of a specified magnitude.

In their paper *Counting sheep: sample size and statistical inference in stable isotope analysis and palaeodietary reconstruction*, Pearson and Grove (2013) carried out statistical simulations to determine the optimal sampling size for a large dataset of sheep (n=174) whose dietary intake was influenced by a combination of anthropogenic and environmental influences. The results showed that when using at least 8 samples per group, the 'true mean' of the group was represented in over 90 % of the cases. When using fewer than 8 samples per group, the 'true mean' was often not captured. Sample sizes larger than 40 presented minimal improvement to the true mean estimation (< 0.3 ‰). While sample sizes between 8–40 samples may prove ideal for estimating group means, it is important to remember that estimation of group variability may require larger sample sizes (Syväranta et al. 2013). Additionally, not all samples may pass quality control assessment and may thus yield unreliable results. For this reason, we recommend sampling slightly above the optimal number of samples (per species or group).

Isotopic values of faunal remains can be used in palaeodietary reconstructions to estimate the underlying ecological 'baseline' values (discussed more in Section 2.6) or to investigate questions pertaining to the management and dietary patterns of the animals themselves (e.g., Cucchi et al. 2016; Nitsch et al. 2017). As with sampling human remains, faunal sampling strategies should consider the benefit of carrying out power analyses that capture estimates of possible sources and magnitudes of variability. Since it cannot be assumed that animals from spatially or chronologically distinct horizons had the same diets, it is advisable to carry out test runs to assess the variability in the isotopic values of fauna from distinct contexts or levels. If planning to reconstruct dietary food-webs using isotope mixing models (e.g. Parnell et al. 2013; Fernandes 2016), careful consideration should be given to identifying and quantifying possible sources of

variability within the consumers, such as intra-individual variability in biological tissues and inter-individual differences in metabolic rates (Cheung and Szpak 2021; Hyland et al. 2021). Additionally, variability within the food items themselves (caused by, e.g., manuring, cooking and fermenting; Royer et al. 2017) and limitations posed by an incomplete record of possible food items (where some may have not preserved in the archaeological record) should also inform sample selection and data interpretation. For detailed advice on sampling and using dietary mixing models, we suggest that researchers follow the best practices outlined in Cheung and Szpak (2021).

Unequal representation of distinct elements in skeletal assemblages often poses a limitation to identifying the desired Minimum Number of Individuals (MNI) from each archaeological context of interest. When skeletal assemblages are large (e.g., the faunal assemblage at Neolithic Makriyalos, northern Greece), the issue of sampling unique individuals more than once can be circumvented by selecting bones of a single element and body side for each taxon. At Makriyalos, Paul Halstead and his zooarchaeological team determined that selecting *humeri* from domestic herbivores and *radii/tibiae* from red deer would provide the largest MNI for each taxon (Vaiglova et al. 2018). At sites with smaller/less well-preserved bone assemblages (e.g., Kouphovouno, Vaiglova et al. 2020a), it may not be possible to only sample one element. In this instance, the researchers should make the best effort to minimize the possibility of sampling an individual more than once while achieving the coverage necessary to address the research question. In both scenarios, including information about the samples (element, side, element portion, and other relevant information, e.g., Balasse et al. 2013) can help demonstrate the researchers' confidence in having sampled distinct individuals.

Preservation of bones is determined by the ability of the organic and inorganic phases to survive in distinct climatic and pedogenic contexts. The organic phase is primarily composed of collagen, a protein that can remain intact in archaeological contexts up to 100,000 years old (Jones et al. 2001), unless an archaeological site is located in the tropics, or in regions with very hot/dry summers and wet winters. Because bone collagen enables measurement of both stable carbon and nitrogen isotope values, it has become the preferred sampling material for palaeodietary studies; although it primarily provides information about the protein component of consumers' diets (Sullivan and Krueger 1981; Ambrose and Norr 1993; Lee-Thorp 2008).

With its high porosity and low crystallinity, the inorganic bone mineral bioapatite can become significantly altered through long-term interaction with soil (Lee-Thorp 2008; Clementz 2012). The surface area of bone is amenable to bicarbonate adsorption (Koch et al. 1997), which can likely alter the stable carbon ($\delta^{13}C$) and oxygen ($\delta^{18}O$) isotope composition of bone apatite. Nevertheless, bone bioapatite can provide reliable stable isotopic measurements when it follows "a pathway to preservation, rather than to destruction" (Lee-Thorp and Sponheimer 2003, p. 210). This can happen in dry burial environments which favor slow removal of collagen with relatively fast increase in crystallinity leading to million-year-old bone resembling enamel in its crystallographic features (Lee-Thorp and Sponheimer 2003). Thus, preservational context plays a role in the potential of bone mineral to retain its *in vivo* stable isotopic composition and should be carefully considered when identifying bones suitable for isotopic analysis.

Strontium, a mobile trace element in soils and sediment, can easily substitute for calcium in bone bioapatite to the point where it becomes impossible to distinguish between *in vivo* strontium and contamination from soil. Experiments have shown that bones that are calcined (i.e. heated to temperatures that turn them white) can preserve their *in vivo* strontium isotope ratios, $^{87}Sr/^{86}Sr$, (Snoeck et al. 2015), but not their $\delta^{13}C$ or $\delta^{18}O$ values (Snoeck et al. 2016). For this reason, while calcined bone may provide reliable samples for strontium isotope ratio measurements in studies of residential mobility, non-calcined bone should be avoided or used only to determine local baseline $^{87}Sr/^{86}Sr$ ratios.

2.2 Tooth enamel

Tooth enamel is commonly analyzed using varying scales of resolution. 'Bulk' tooth samples (i.e. one measurement per individual; e.g., Roberts et al. 2015) approximate the average value of all isotopic inputs over the course of tooth formation. This can provide a way to compare multi-seasonal averages between individuals. Similarly to bone collagen sampling discussed above, power analyses can be useful for estimating the optimal number of samples required to answer the research question (Pearson and Grove 2013; Syväranta et al. 2013; Cheung and Szpak 2021).

Incremental sampling that follows the direction of tooth growth provides a sequence of samples that can be used to assess sub-annual changes in isotopic inputs throughout tooth formation (Balasse 2003). The temporal resolution of these sequences varies depending on the sampling technique. Manually sampled sequences (extracting 1mm–wide samples spaced ~1mm apart, e.g., Balasse 2012; Janzen et al. 2020) provide attenuated values averaged over several months of tooth formation (Balasse 2003; Zazzo et al. 2005; Pederzani and Britton 2019). Micro-driling with laser-ablation mechanisms (e.g., Copeland et al. 2008; Pryor et al. 2020) provide higher resolution time slices averaged over a few months, while targeted micro-drilling using histological maps of developmental features (i.e. daily growth lines) can approximate isotopic inputs with weekly resolution (e.g., Austin et al. 2016; Smith et al. 2018). Due to budgetary constraints, incremental sampling often necessitates analysis of fewer individuals than bulk sampling. For example, while Roberts et al. (2015) were able to analyze 300 individuals with bulk sampling, incremental sampling was more restrictive: Janzen et al. (2020) analyzed 60 individuals, Pryor et al. (2020) analyzed 16 individuals and Smith et al. (2018) analyzed 8 individuals. For this reason, incremental tooth enamel sampling is most amenable for answering questions of individual or group, rather than population-wide, life histories.

A major consideration of sample selection is the likelihood that a tooth preserves a complete annual sequence representative of the full range of variability in seasonal isotopic inputs. Partial sequences from highly worn teeth may represent incomplete annual cycles, and the absence of 'real' maximum or minimum values may preclude the ability to determine which seasons are represented. For example, of the 32 teeth drilled from Byzantine sites in the Negev desert (Vaiglova et al. 2020b), 10 teeth provided incomplete sequences, which are shown in Supplementary Material 2. Without being able to determine which seasons are represented, inclusion of these partial sequences in the analyses would lead to skewed results (e.g., over-representation of summer values and under-representation of winter values). To overcome this problem, only teeth with full annual sequences should be included in inter-teeth comparisons. If it is possible to determine which seasons are represented (e.g., by identifying a clear sinusoidal 'peak' or 'trough'), partial sequences may be used to compare seasonal values (e.g., compare only summer values across individuals). Teeth that complete mineralization over several annual cycles (e.g., horse teeth, Hoppe et al. 2004: Table 2) may also provide skewed results if teeth of varying

wear stages record formation over varying numbers of annual cycles. To overcome this issue, fixed lengths of teeth crowns (e.g., 3cm for horse teeth, Britton et al. 2019) can be drilled to keep the temporal resolution as constant as possible.

The choice of tooth is constrained by the timing of tooth mineralization and the ability of different teeth to answer the research question. Unlike bones, which remodel during the lifetime of individuals, the elemental composition of teeth remains generally constant once formation is complete (Pederzani and Britton 2019). The precise timing of tooth development is primarily determined by species and tooth type. Earlier forming teeth (e.g., first molars, M1) can provide information about initial life patterns such as cessation of nursing, while later forming teeth (e.g., third molars, M3) record dietary inputs in later years (Weinreb 1964; Balasse et al. 2001; Hillson 2005). The duration of mineralization also determines if the tooth will record full annual sequences or not. For example, due to their shorter lengths and faster mineralization schedules, pig molars do not record full annual sequences and are thus not the tooth of choice for studying pig seasonal dietary and mobility patterns. These objectives can instead be accomplished by incrementally sampling their longer-forming incisors (Frémondeau et al. 2012; Zhang et al. 2021).

Unlike bone apatite, tooth enamel is relatively robust to diagenesis (Zazzo et al. 2004; Lee-Thorp 2008). Its higher crystallinity prevents exogenous trace elements like strontium from being incorporated deep inside the mineral structure (Macfadden and Cerling 1996; Snoeck et al. 2015). More recent research, however, has revealed that tooth enamel is not always as structurally robust as we routinely assume (Wood et al. 2021; Weber et al. 2021). Quality criteria outlined in Table 1 can help assess the integrity of tooth enamel for isotopic analysis prior to analysis of the results. For high-resolution analyses that employ instruments such as a Laser Ablation Inductively Coupled Plasma Mass Spectrometer (e.g., Sponheimer et al. 2006), or Secondary High Resolution Ion Microprobe (e.g., Aubert et al. 2012), elemental mapping of trace elements like uranium can be used to identify portions of enamel that may have undergone diagenetic alteration. This information can help target regions that are most likely to provide *in vivo* composition.

Depending on the questions, researchers working with tooth enamel may choose to analyze either $\delta^{13}C$ and $\delta^{18}O$ values of structural carbonate or $\delta^{18}O$ values of enamel phosphate. Enamel

phosphate is more resistant to diagenetic alteration (Zazzo et al. 2004), and is the preferred material for addressing questions of palaeoclimate. On the other hand, matching $\delta^{18}O$ and $\delta^{13}C$ values of structural carbonate makes these samples amenable to addressing questions of both seasonal mobility and dietary intake (Makarewicz and Pederzani 2017; Pederzani and Britton 2019).

2.3 Charred plant remains

Sample selection for plant isotope analysis is also constrained by both the research question and the depositional history of the assemblage. Seeds from primary *in situ* contexts (e.g., storage deposits) can provide information about growing conditions on single years or locations (e.g., Styring et al. 2016). Seeds derived from secondary contexts necessarily provide an averaged indicator of cultivation conditions that can span multiple years or locations (e.g., Vaiglova et al. 2020a). Seed samples can either be obtained in bulk (i.e. several seeds from an individual archaeological context) or as individual single-seed samples, and the choice depends on the research aims. Single seed analysis can be useful for assessing inter-crop variability within sampling contexts, and to avoid the possibility of averaging crop growing conditions across different cultivation events. However, this analysis may not be feasible in cases where the amount of nitrogen (% N) of small samples is too low for reliable analysis of stable nitrogen isotope values ($\delta^{15}N$) (Szpak and Chiou 2019).

Bulk sampling is effective if the seeds derive from primary contexts where it can be assumed that the crops represented were cultivated in the same year; and this only applies to species of annual plants. It can also be an appropriate sampling strategy if the assessment of inter-plant variability is not crucial answering the research question. Experimental research on modern crops from a single growing season and treatment condition suggests that analysis of homogenized bulk samples of ~10 cereal grains register variability of ±0.5 ‰ in $\delta^{13}C$ values and ± 1.0 ‰ in $\delta^{15}N$ values, while analysis of smaller numbers of grains (i.e., < 5) show much greater variability (Nitsch et al. 2015). Therefore, to assess the variability between growing conditions represented in single *in situ* deposits or storage contexts, we recommend that researchers aim to obtain several (ideally more than 3) bulk samples per context.

The most common way seeds are preserved in the archaeological record is via charring, although uncharred or lightly toasted seeds have been found in rare waterlogged, frozen, or desiccated contexts. Outside of these rare preservation conditions, seeds need to be exposed to temperatures above ~215 ºC to become highly resistant to microbial attack and survive long-term burial. Seeds charred above ~400 ºC become highly distorted and unamenable to species identification (Charles et al. 2015). Because of their highly porous structure, charred material can absorb high amounts of strontium from soil and reliable application of $^{87}Sr/^{86}Sr$ ratio analysis for provenancing archaeological plant material is currently highly problematic (Styring et al. 2019).

For analysis of stable carbon and nitrogen isotope values, selection of charred seeds requires the ability to assess the effects that charring has had on the *in vivo* isotopic composition of the material. Experimental studies have shown that temperature, and to a lesser degree duration, alter both $\delta^{13}C$ and $\delta^{15}N$ values of cereal grains and pulse seeds (Nitsch et al. 2015; Stroud et al. in prep). These experiments have also shown that the isotopic values of seeds charred between 215–300 ºC are altered in ways that can be modelled and corrected (Nitsch et al. 2015; Stroud et al. in prep). Ascertaining the temperature at which seeds have been charred requires visual inspection of both the surface morphology and the internal seed structure (cf Charles et al. 2015, Stroud et al. in prep). Assessment of the internal matrix is crucial, as recent experiments have shown that even cereal grains whose surface morphological characteristics suggest that they are 'ideally' preserved can be completely void on the inside (Stroud et al. in prep). Because the preservation potential of cereal grains can vary within and between archaeological contexts, plant material should be assessed on a grain-by-grain basis to ensure that only grains that are sufficiently preserved are measured and interpreted (Stroud et al. in prep). To demonstrate confidence in sample reliability, we suggest that researchers publishing plant stable isotope results provide access (in supplementary materials or online databases) to photographs of each analyzed seed that is included in the analysis.

The isotopic and morphological changes caused by charring are currently only understood for a relatively small number of cereals: emmer (*Triticum dicoccum* (Schrank) Schübl.), einkorn (*Triticum monococcum* L.), hulled barley (*Hordeum vulgare var distichum* L.), bread wheat (*Triticum aestivum* L.), spelt (*Triticum spelta* L.), oat (*Avena sativa* L.), rye (*Secale cereale* L.), maize (*Zea Mays* L.), and one millet species: pearl millet (*Pennisetum glaucum* L.); and two pulse

species: lentil (*Lens culinaris* Medik.) and pea (*Pisum sativum* L.) (Nitsch et al. 2015; Charles et al. 2015; Styring et al. 2019; Hart and Feranec 2020; Stroud et al. in prep). For this reason, before further charring experiments are carried out, caution should be taken when interpreting isotopic data derived from species for which the effect of charring is unknown.

2.4 Molluscan shell and eggshell

Shells of aquatic and terrestrial molluscs preserve valuable short-term and incremental records of palaeoenvironmental parameters such as water temperature and C_3/C_4 plant abundance (Goodfriend 1992; Johnson et al. 1998; Andrus 2011). However, aspects of the animal's life history, habitat, and physiology may incorporate systematic biases and offsets into the preserved environmental record (Schöne 2008; Ivany 2012). An animal typically slows down or speeds up shell production under particular environmental conditions (e.g., temperature, food availability) or physiological circumstances (e.g., reproductive cycle, recovery from predatory attacks), resulting in those periods becoming either over- or underrepresented in their internal record. Additionally, an animal's behavior may change over the course of its life. For example, many marine molluscs will move to higher or lower regions within the intertidal zone as they grow larger (and thus record a changing temperature range across their shell), or their growth rate will slow down with age (Schöne 2008). As these factors are unique to each species, it is important to carefully consider the ecology and biology of the given species. A robust study of shells from modern and well-characterized environments can help to assess the sampling resolution required to attain the desired temporal resolution of the temperature record (e.g., daily, monthly, annual), as well as to evaluate any systematic biases (e.g., seasonal, ontogenetic). It can also provide information on any inter-individual variation caused by differences in life histories and microhabitats, which should be taken into consideration when deciding how to apportion samples within individuals or entire assemblages (e.g., Galimberti et al. 2017).

In addition to validating the usefulness of particular species for stable isotope studies, shell samples need to be assessed for preservation (Leng and Lewis 2014). Carbonate shells are prone to dissolution and recrystallisation in many depositional settings where water is present in the sediments. Unfortunately, the process of recrystallisation may overwrite the original environmental signal recorded in their $\delta^{18}O$ and $\delta^{13}C$ values, and these shells should be regarded

as unsuitable for stable isotope analysis (Brand 1989). Additionally, cooking practices may have also altered the shell structure, such that burned or heated shells must likewise be avoided (Milano et al. 2016). For these reasons, it is necessary to evaluate the crystal microstructure of individual shells, especially where preservation varies considerably across a site.

Visual examination of sectioned shells under a microscope can quickly reveal changes to shell structure, with recrystallized shell looking amorphous and 'flaky', rather than highly crystalline. Many aquatic species and almost all terrestrial species precipitate part or all of their shell in the aragonitic polymorph of carbonate, as opposed to calcite (Lowenstam 1954; Goodfriend 1992). As recrystallized shell is always calcitic, and the two polymorphs are simple to distinguish with a range of analytical chemistry methods (i.e. XRD, Raman spectroscopy, chemical staining), assessment of the presence of aragonite and calcite offers a convenient diagenetic check during sample selection. The methods, however, vary considerably in terms of how destructive and/or quantitative they are. FTIR (Fourier Transform Infrared Spectrometry) with an ATR (Attenuated Total Reflectance) attachment offers a suitable balance of analytical speed and simplicity, damage to the sample, and polymorph quantification (Loftus et al. 2015). This enables each powdered sample to be assessed for diagenesis prior to stable isotope analysis.

2.5 Samples obtained during AMS radiocarbon dating

In radiocarbon dating laboratories, $\delta^{13}C$ values are measured alongside radiocarbon determinations for several reasons: 1) to correct for reservoir effects which may have shifted the measured age away from the true age (e.g., marine reservoir effect, Knipper et al. 2020), 2) to correct for fractionation to the ^{14}C:^{12}C ratio that occurred during the graphitization stage of sample preparation, and 3) to systematically normalize all radiocarbon measurements to $\delta^{13}C = –25.0$ ‰, even if they have not been influenced by reservoir or graphitization fractionation effects (Stuiver and Polach 1977; Reimer et al. 2004). These measurements are performed either directly on the Accelerator Mass Spectrometer (AMS) or by directing a portion of the CO_2 gas from the AMS to the Isotope Ratio Mass Spectrometer (IRMS). However, these measurements are often not usable for palaeodietary/palaeoclimatic interpretations, or for identifying small differences between samples (e.g., < 0.2 ‰) for three reasons (personal communication: Richard Staff, Scottish

Universities Environmental Research Centre; David Chivall and Peter Ditchfield, Oxford Radiocarbon Accelerator Unit).

First, the graphitization stage during sample preparation induces fractionation, which essentially alters the ^{13}C:^{12}C ratio by an indeterminate amount. Secondly, much larger samples are injected into the AMS than would be suitable for IRMS and this can push the IRMS to its limit, potentially introducing error. Third, the AMS $\delta^{13}C$ values are often normalized using a one-point rather than two- or multi-point calibration, which leads to lower measurement accuracy. At the Oxford Radiocarbon Accelerator Unit, the uncertainties attached to AMS $\delta^{13}C$ measurements are not calculated, but the system flags measurements only when the $\delta^{13}C$ value obtained during combustion for graphitization and the $\delta^{13}C$ value obtained by the AMS differ by > 5.0 ‰. For this reason, when clients request accompanying $\delta^{13}C$ values, separate aliquots of the samples are prepared for analysis using an IRMS, where the uncertainty is often better than 0.5 ‰. Altogether, stable carbon isotope values produced via AMS dating are not the same as those obtained on an IRMS. Researchers should thus only use AMS $\delta^{13}C$ values for palaeodietary/palaeoclimatic research when their accuracy and precision can be determined and when the measurement error fits within the expected uncertainty for answering the research question.

2.6 Final word on sampling: disentangling natural from anthropogenic factors

The extent to which isotopic measurements of archaeological materials can shed light on questions of archaeological interest is constrained by our ability to isolate the effects of human behavior from the effects of environmental factors that may have simultaneously influenced the chemical composition of the sample material. For interpreting strontium isotope ratios, comparative maps of biologically available strontium can be built by modelling the spatial distributions of $^{87}Sr/^{86}Sr$ ratios of modern plants, surface water, snail shells, insects (Evans et al. 2010; Snoeck et al. 2020; Wang et al. 2021; Holt et al. 2021). For interpretation of archaeological stable carbon and nitrogen isotope values, the ecological background can be approximated through measurement of specimens with low likelihood of impact from anthropogenic factors (e.g., wild species, unmanaged landscapes). However, the identification of non-anthropogenically influenced samples is fraught with many challenges.

Firstly, rather than focusing on estimating average base*line*, we should instead focus on approximating a range (or 'base*interval*') of local background values. Baseintervals will be able to capture both the natural variability within the given ecology and the measurement error accompanying the isotopic measurement. Secondly, baseintervals should not be thought of as static. Since isotope ecologies are determined by a range dynamic factors including precipitation patterns, soil dynamics, and distribution of local vegetation and faunal species, they may have changed on annual, decadal, centennial, or millennial scales (Gannes et al. 1997). These changes would have caused reverberating effects through the food chain and influenced the isotopic composition of food sources (primary and secondary producers) as well as consumers in distinct trophic levels. Measurements of reference samples that are as contemporary as possible to the samples of interest are therefore key for demonstrating that observed patterns in isotopic data are most likely to be the result of changes in human behavior rather than shifts in the underlying ecology.

Thirdly, multiple plant and animal species may be necessary to capture the isotopic variability of specific vegetation communities, ungulate dietary niches, or human food production systems (Codron et al. 2011, 2013). This is where questions may arise regarding whether or not the comparative species may have been impacted by anthropogenic factors, especially in archaeological contexts where intensive agriculture or landscape management was pursued. For example, wild herbivores may have raided crop fields, and their bone collagen $\delta^{13}C$ and $\delta^{15}N$ values would thus reflect the consumption of both natural forage and plants that may have been influenced by manuring, irrigation or crop selection (e.g., Gillis et al. 2020). Human activities such as woodland clearance may have forced wild herbivores to graze in open environments with higher plant $\delta^{13}C$ values compared to those in closed–canopy woodlands (Bonafini et al. 2013). Additionally, weeds may have grown in arable cultivation plots while nut or fruit trees may have been planted on the edges of managed fields, where they may have been similarly influenced by human activities that impact the $\delta^{13}C$ and $\delta^{15}N$ composition of soil.

To identify faunal species most appropriate for building reference baseintervals, it is important to consider the specific dietary adaptations (e.g., browsing vs. grazing) and physiological requirements (e.g., obligatory vs. non-obligatory drinking) of the available species. Additional

considerations include questions of population size, forage availability, and predation pressure, as these may have influenced dietary niche size and composition and caused directional isotopic changes to wild animal populations over time (Codron et al. 2011).

Identifying comparative plant species needs to take into account both the likelihood that they grew in managed soils, as well as possible isotopic variability between distinct plant parts and growing seasons (Tieszen 1991; Fraser et al. 2011; Cernusak et al. 2013). Although the directional effects of certain crop management systems (e.g. application of farmyard manure or other organic waste, small or large-scale systems of water management) have been demonstrated experimentally (Bogaard et al. 2007; Fraser et al. 2011; Wallace et al. 2013; Flohr et al. 2019; Jones et al. 2021), the magnitude of the effects would be challenging to estimate in environments where experiments have not been carried out. For example, using the manuring offset of crops that were grown in Britain, Germany, and Romania (Fraser et al. 2011) to estimate the degree of manure application on crops grown in South Africa would be inappropriate, because the two environments are not directly comparable due to differing climate, soil types, and preservation potential of archaeological materials.

With these caveats in mind, sampling for comparative baseintervals should aim to capture, to the best of a researcher's ability, the full range of isotopic variability within a given ecosystem. Interpretations of measured archaeological values should then be constrained by predictions surrounding the extent to which the measured species may have been influenced by natural and anthropogenic factors (Roberts et al. 2017b; Knipper et al. 2020; Gillis et al. 2021). Additional lines of evidence, such as pollen records, patterns of agro-pastoral intensification and extensification (Styring et al. 2017; Price et al. 2020; Marston 2021), human landscape modification (Rull et al. 2021), and local-scale environmental change (Roffet-Salque et al. 2018) can furthermore be used to clarify the mechanisms that may have influenced the isotope composition of archaeological flora, fauna, and human remains.

3. Using exploratory data analysis to guide formal analysis

Exploratory data analysis is an important part of any analysis and should be carried out prior to formal analysis through inspection of graphical representation of the data. Simple univariate plots

such as histograms, boxplots, and violin plots can shed light on the behavior of individual variables, while scatterplots can elucidate the connections between them. Exploratory data analysis should not be a rote exercise, but should instead involve a thoughtful investigation of the data. Many books have been written about the principles of good data visualization (e.g., Cleveland 1994; Wilkinson 2005) and packages such as ggplot2 in R (Wickham 2016) provide a variety of flexible tools for effectively exploring data.

3.1 Suggestions for effective data visualization

Simple aesthetic adjustments and choices can have a large impact on the readability of figures. One does not need to be proficient with coding nor to use programs like R and Python to produce graphs that are aesthetically pleasing and easily readable. Here we outline some of the most relevant advice for visualization of isotopic data; referenced examples are reproduced in Supplementary Material (SM) 1.

(a) Plot the data in ways appropriate for addressing the research question. If the objective is to assess how much variation there is between individual data points, standard deviation is most effective for showing the spread of the data. If the objective is to assess how good the estimation of the group mean is, standard error is more appropriate. Boxplots, either individual or plotted side by side to show group comparisons, can highlight important and interesting patterns in datasets. For archaeological data which may be non-normally distributed (bimodal, skewed), violin plots may be more informative than boxplots (e.g., Cavazzuti et al. 2019, Fig.6, in SM 1.A). Bivariate scatterplots (e.g., Levin et al. 2011, Fig.8, in SM 1.B), especially when enhanced by scatterplot smoothers (such as loess curves, e.g., Wierzbowski et al. 2017, Fig.8, in SM 1.C), are an effective way of exploring the relationships among variables. The idea can, with care, be extended to rotating three-dimensional plots that may reveal relationships that are not evident from collections of bivariate plots alone (see, for example, Wainer 2007, pp. 126–127).

Crucially, the most appropriate method of visualization will depend on the relationships being explored. For example, with intra-tooth sequences of isotopic values, one can look at the *shape* of the sequences to gather information about seasonal variability in dietary inputs (e.g., Vaiglova et al. 2018, Fig.10, in SM 1.D), the *intra-tooth variability* to assess the magnitude of variation in

dietary inputs between different individuals (e.g., Vaiglova et al. 2020b, Fig.2, in SM 1.E), or the *position* of different parameters (such as the maximum $\delta^{18}O$ value) on the tooth axis (e.g., Balasse et al. 2021, Fig.2, in SM 1.F). Figures can thus be used appropriately to highlight these relationships of interest. See Weissgerber et al. (2019) for advice and examples of matching visualization to research question.

(b) Do not use randomly assigned combinations of symbols and colors. Looking at the spread of data to see which species, groups, or individuals overlap allows researchers to choose symbols based on what will make the data points most easily discernible. The size of plotting symbols can often be controlled (made smaller or larger), while color can be shaded and made more or less transparent. This can be used to the researcher's advantage, yielding more flexibility beyond the software defaults. For example, in the dataset presented in Vaiglova et al. (2021, Fig.S2, in SM 1.G), many human datapoints overlapped with those of dogs and pigs. Thus, the authors chose to use different colors, rather than symbols, to represent these three species. Where less overlap was visible (between humans and cattle/sheep/goats), different symbols were used. Altogether, only 5 colors were used to represent 18 species, and the color-symbol combinations were used in ways that led to least overlap between similar symbols.

(c) Keep graphs as simple as possible. Overcrowding figures with too many overlapping groups may obscure actual patterns of interest from being discernible. For example, if individuals from distinct age or sex categories have meaningfully different isotopic values, it is useful to display them separately. However, if their values are not meaningfully different, they can be combined into fewer groups to prevent the reader's eye from trying to detect patterns where they do not exist. Where a large number of units are chosen to be displayed, an effective strategy is to group them into clusters that share the same symbol, and that are further differentiated with different colors (e.g., Knipper et al. 2017, Fig.3, in SM 1.H)

(d) Make graphs readable to people with color vision deficiency. Millions of people around the world suffer from color vision deficiency (CVD), with prevalence for just one type of deficiency (red-green) reported to be between 4–8% in men and ~0.4% in women (Birch 2012). To make publication figures as readable as possible to the wider audience, we suggest that researchers use

CVD-friendly palettes to prepare their graphical images, and run them through a simulator (e.g., https://www.color-blindness.com/coblis-color-blindness-simulator/) to check that they are compatible with different types of deficiencies. If CVD-friendly palettes are not used, using black borders around symbols (as opposed to colored borders) makes the colors more discernible. The combination of green and red is most commonly challenging, but different shades of purples and oranges can work well with the full range of color deficiencies. Vaiglova et al. (2021, in SM 1.G) used the IBM palette to make patterns in a scatter plot of δ^{13}C vs. δ^{15}N values visible on a yellow-shaded background. If combinations like green and red are used, the groups can be distinguished using different symbols (e.g., Levin et al. 2011, Fig.7, in SM 1.B), but this only works with small numbers of groups.

(e) Use legends to your advantage. Many researchers prefer to spell out the graphical parameters of figures using caption descriptions such as "Samples from Area A are shown with closed diamonds (◆, solid regression line), samples from Area B are shown with open diamonds (◇, dashed regression line)". We argue that it is considerably easier for readers to discern the aesthetic assignments of graphs by looking at legends rather than reading long descriptions in captions. Vaiglova et al. (2020b, Fig.2, in SM 1.E) provide an example of how nested legends can be used to describe the aesthetic assignments of several variables (species, chronological period, site, material measured) for separate panels, and for all panels combined.

4. Formal analysis: do not focus on statistical significances

For decades, statisticians and researchers in many fields of science have decried the rote application of null hypothesis testing, specifically the aspect of it that compares a computed *p*-value to an arbitrary threshold (often 0.05 or 0.1) and declares the result "statistically significant" or not. The *p*-value itself is prone to many misconceptions (see for example Goodman 2008; Greenland et al. 2013), while statistical significance is often confused with practical significance (i.e., those results that are of scientific importance). This way of thinking has become entrenched to the point that publication can hinge on whether or not results achieve statistical significance, in turn opening the door to various questionable research practices such as *p*-hacking (Simmons et al. 2011; Head et al. 2015).

In 2016, the American Statistical Association (ASA) responded to these worrying and widespread misuses of statistical testing by issuing a 'Statement on *p*-values' (Wasserstein and Lazar, 2016). The statement provides important clarifications on what *p*-values do and do not measure, what they do and do not represent, and what they should and should not be used for. The statement includes six principles; commentaries from the drafters of the statement (not all of whom agreed with the final version) are available as online supplements. Subsequently, in 2019, *The American Statistician* published a Special Issue on moving to a world beyond '$p<0.05$' with articles espousing a range of views on the controversies now swirling around this most common of statistical tools. The lead editorial to the Special Issue (Wasserstein et al. 2019) called for the abandonment of strict thresholding of statistical quantities (such as the *p*-value) as well as of the use of the term "statistically significant."

Drawing on these debates, this section aims to provide some of the most pressing clarifications on the uses and misuses of statistical significance testing in isotope archaeology. It outlines advice on using the most common statistical test applied to isotopic data: comparing meaningful differences between groups (e.g., human groups distinguished by age, sex, social status, type of burial, or between different plant or animal species). While we do not wish to advocate for the abolishment of *p*-values, we strongly reiterate the statistical community's advice against 1) using the term 'statistically significant difference' (and related practices such as using boldface or asterisks in a table to highlight specific groups) and 2) employing 'significance thresholds' (such as 0.05 and 0.1), which can inaccurately distinguish between *p*-values of 0.051 and 0.049 (values that are, in fact, not meaningfully different).

4.1 How do we test for meaningful differences between groups?

In a study investigating ancient agricultural management at the Neolithic site of Kouphovouno, southern Greece, Vaiglova et al. (2020a, p. 43) assessed differences in δ^{13}C values between four main crop species (free-threshing wheat, hulled barley, einkorn and pea). The paper reported:

The data for the four main species (where n > 1) are normally distributed (Shapiro-Wilk test, W = 0.95, p = 0.09) and homogenous (Levene's test, F(3,38) = 1.12, p = 0.35), so an ANOVA test was used to assess statistically significant differences between the mean δ^{13}C values of the four groups. The results indicate that there are significant differences (F(3,38) = 25.26, p < 0.01) and

a post hoc Bonferroni test reveals that the differences are between free-threshing wheat and hulled barley ($p < 0.01$), free-threshing wheat and pea ($p < 0.01$) and einkorn and pea ($p = 0.004$).

While this example does conform to many current conventions on statistical reporting, it is incomplete in important ways. For example, the actual means and their standard errors are missing. Instead, the authors provide only test statistics and (sometimes thresholded) *p*-values. This leaves the reader unable to judge if the differences among the groups are of scientific interest and deserving of further exploration; or what magnitude of difference would even be meaningful in this context. Tests for normality and homogeneity (especially the former) are popular in many fields. Strictly speaking, they may not be necessary, since the *t*-test and analysis of variance are relatively robust to minor departures from assumptions under mild conditions (Miller 1986). Furthermore, even if preplanned as part of the analysis, these tests would still need to be included in any corrections for multiple testing (to preserve overall Type I error control), thereby detracting from the comparisons of main interest. And, of course, each decision made on the basis of a hypothesis test is an opportunity for additional errors.

Rather than focusing on the test statistics and their *p*-values, along with accompanying statements about statistical significance, a more appropriate approach would be to provide plots showing the data distribution in each group along with the point estimates (i.e. mean differences) and their associated confidence intervals (CI). Instead of focusing on whether or not these intervals include zero, the researchers should define and focus on the range of "compatible" values by stating, for example, what differences would be meaningful to interpret. In this way, a richer engagement with the data can be presented, replacing the dry and arbitrary assessments of statistical significance, or lack thereof.

Following this advice, a more accurate summary of the analysis of the differences in δ^{13}C values between the crops at Kouphovouno is as follows:

Figure 1 shows the distribution of the δ^{13}C values of the four main crops in the assemblage, where mean δ^{13}C values are –24.3 ± 0.6 ‰ (n=15) for barley; –22.7 ± 0.5 ‰ (n=17) for free-threshing wheat; –23.6 ± 0.8 ‰ (n=6) for einkorn; –25.0 ± 0.9 ‰ (n=7) for pea. Small sample sizes make it difficult to assess normality, however, the crops exhibit somewhat similar levels of variability. A

one-way analysis of variance was conducted to compare the $\delta^{13}C$ values of the four crops; here, F(3,38)=25.26, with a p-value of 0.002, indicates that there are differences in the mean $\delta^{13}C$ values between the crop species.

To explore these differences further, Tukey's HSD intervals were computed to control for post-hoc multiple comparisons. Results are shown in Figure 2 and Table 2. Pea appears to have lower values than free-threshing wheat, einkorn, and hulled barley. However, inspection of the confidence intervals in the comparison with hulled barley indicates that slightly larger values are also compatible with the collected data and the small number of observations in einkorn make it hard to draw conclusions. Hulled barley has lower values than free-threshing wheat and einkorn, though again the confidence intervals in the latter indicates that slightly larger values are also compatible with the collected data. Finally, free-threshing wheat has higher values than einkorn, although with the small number of observations of these groups, it is hard to be definitive.

Focusing on isotopically meaningful differences—which we define as > 1.0 ‰ in this context—and keeping in mind the small sample sizes, the distinctions between the two types of wheat (free-threshing wheat and einkorn) and peas can be explained by peas' higher demand for water during the seed-filling period. As wheat and barley grown under the same watering conditions exhibit physiologically-determined differences (with barley having $\delta^{13}C$ values c. 1–2 ‰ lower), the measured difference of –1.6 ‰ (CI: –2.3, –0.8) is compatible with the two crops growing under the same watering regime.

4.2 Further comments on reporting meaningful differences: decimal places and effect sizes

When reporting *p*-values and other values such as means or test statistics, the question often arises as to how precisely those results should be reported. For *p*-values, two or three digits after the decimal are usually sufficient, hence 0.049 is preferred to 0.049326. Little important knowledge is gained by the extra trailing digits. On the contrary, unnecessary digits may impart a false sense of precision. Likewise, there is very little to be learned by reporting a *p*-value of 10^{-16}, or other similarly small numbers, which are in essence artifacts of machine precision. For averages and test statistics, a similar logic applies, and we should refrain from implying more precision than is supported by the data themselves. Ehrenberg (1981) advocates rounding to "two effective digits" to balance the lightening in cognitive load and information loss. In addition, meaningful isotopic differences may not necessarily lie at the highest level of measurement precision, and sample

homogeneity needs to be taken into account when deciding how many decimal places to report and interpret (see more in Section 5.3).

It is good habit to report effect sizes or other meaningful summaries of the data, with their accompanying degrees of freedom, in addition to test statistics. These should have measures of uncertainty attached to them (e.g., standard errors or standard deviations as appropriate). If the data have been transformed for analytical purposes, it can be helpful to bring the results back to the original scale for interpretation. As in the rewritten example above, it is advisable to discuss the effect sizes and the uncertainty in the context of the original data and the scientific questions, rather than simply stating differences to be 'statistically significant'.

5. Writing scientifically rigorous publications

5.1 Avoiding common misuses of terminology

To ensure that isotope research is easily comprehensible to practitioners working across varying subfields of isotope geochemistry, it is important that the data be disseminated using widely accepted conventions and standardized definitions of terms. Coplen (2011) presents the guidelines for reporting isotope ratio measurements established by the Commission on Isotopic Abundances and Atomic Weights of the International Union of Pure and Applied Chemistry and Roberts et al. (2017a) provides examples of the most common misuses of the terminology in archaeological research. Here we summarize some of the most misused conventions:

1. Isotope 'delta' notation (e.g., $\delta^{13}C$, $\delta^{15}N$, $\delta^{18}O$) is defined as the *relative difference of isotope ratios* between samples and international standards. It is in effect a 'ratio of a ratio' and should by convention be referred to as a 'value' (and not a 'ratio', 'signature', 'composition' or 'fingerprint'). Notation that expresses the relative abundance of specific isotopes (e.g., $^{87}Sr/^{86}Sr$) should be referred to as a 'ratio' (and not a 'value', 'signature' or 'fingerprint').
2. The words 'enriched'/'higher' and 'lower'/'depleted' are not interchangeable. Stable isotope *values* (e.g., $\delta^{13}C$, $\delta^{15}N$, $\delta^{18}O$) and *ratios* (e.g., $^{87}Sr/^{86}Sr$) can be higher or lower, but they cannot be enriched or depleted. Enrichment and depletion describe the abundance of specific isotopes themselves (e.g., ^{13}C, ^{15}N, ^{18}O) and need to specify which samples are

being compared to one another, e.g., "Sample A is enriched in ^{15}N compared to Sample B".

3. ‰ is permil (parts per thousand), and not equivalent to ppm (parts per million).
4. The terms 'isotope' and 'isotopic' have distinct uses. 'Isotope' is used when the word is modified (e.g., "The stable carbon isotope value") while 'isotopic' is used when it stands alone (e.g., "The isotopic composition" or "isotopic variability") (Sharp 2017).
5. Negative values should be denoted with an en (–) dash and not a hypthen (-) (Coplen 2011).
6. The 'delta' symbol in $\delta^{13}C$, $\delta^{15}N$, $\delta^{18}O$ should be italicized as per the convention of denoting quantity symbols (Coplen 2011).

5.2 Transparency in measurement calibration and error estimation are indispensable for establishing data reliability

Two of the biggest inconsistencies in isotope archaeology are the frequency with which publications report how raw isotope values were converted to calibrated δ-(delta) values (even though the choice of calibration method can have a significant impact on the accuracy of the results; Pollard et al. 2011; Szpak et al. 2017) and the level of detail provided in assessing measurement reliability. Precision and accuracy are two distinct indices of measurement error, and both are important for establishing data reliability. Accuracy refers to the 'trueness' of the measurements (i.e. how close they are to the true values of the samples), while precision captures the reproducibility of the measurements (i.e. whether repeated measurements on one instrument produce the same results). Yet, most publications in isotope archaeology do not report these two indices, nor do they specify what calibration methods (e.g., one-point, two-point, multi-point calibration) were used to calculate the δ-values. In a review of 487 archaeological publications presenting original isotopic data, Szpak et al. (2017) found that 83% of the studies provided no information about how the δ-values were calibrated, 95% of the studies did not mention accuracy and only 4% of the studies distinguished between accuracy and precision. The authors concluded that "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" (Szpak et al. 2017, p. 612).

It is necessary to move away from the current reporting of measurement error that provides only relatively uninformative statements such as "Measurement uncertainty for both $\delta^{13}C$ and $\delta^{15}N$ was < 0.1 ‰". To accurately calculate measurement precision and accuracy, both *calibration standards* and *check standards* should be used consistently in every analytical session and publications should report measurement uncertainty for particular runs, rather than 'rolling averages' for a given laboratory (Szpak et al. 2017). This becomes particularly important as archaeological isotope research moves towards larger meta-analyses, where distinctly different measurement errors might render smaller datasets incomparable. Based on recommendations from the Joint Committee for Guides in Metrology (2008) and a series of other studies (p.614), Szpak et al. (2017, p. 614) present a series of best practices for calibrating raw isotope values and calculating measurement uncertainty. They provide a user-friendly Microsoft Excel tool that can calculate precision, accuracy, and a pooled standard deviation of the two for particular analytical sessions, and offer an example paragraph and Online Supplement File showing how to fully report measurement error for a given project.

5.3 The number of decimal places reported should be consistent with both measurement uncertainty and sample homogeneity

Stable isotope values of unknown samples can only be as precise as the reference materials analyzed alongside them. Thus, if the measurement precision is reported as $\delta^{13}C = 0.1$ ‰, sample $\delta^{13}C$ values should only be reported to one decimal place (Bond and Hobson 2012). However, many archaeological materials do not present the same level of internal homogeneity as the reference materials. For example, in a study of inter- and intra-bone variability (assessed through measurement of multiple skeletal elements from animals representing a wide variety terrestrial, marine, and freshwater species), the mean intra-bone variability was 0.63 ± 0.06 ‰ and the mean inter-individual isotopic variation was 1.45 ± 1.15 ‰ (Hyland et al. 2021). Thus, even though precise $\delta^{13}C$ and $\delta^{15}N$ measurements can be obtained to two decimal places, the second decimal place may not provide any meaningful information. Authors should thus make sure that the reported values are consistent across a publication and that the reported error does not give a false impression about the interpretability of the data. In addition, elemental compositions (%C, %N) should be reported to one decimal place, while C:N ratios should be reported to two decimal places (see discussion in Guiry and Szpak 2021).

804

805 Strontium isotope measurements can be precise up to 5 or 6 decimal places (depending on the
806 instrumentation) but in a study investigating intra-individual variation in $^{87}Sr/^{86}Sr$ ratios in human
807 tooth enamel, Knudson et al. (2016) found that differences between individual measurements
808 ranged from below detection level to 0.00015. They thus concluded that meaningful archaeological
809 interpretations about palaeomobility using $^{87}Sr/^{86}Sr$ ratios can be made at the third or fourth
810 decimal place.

811

812 **5.4 Writing with clarity and brevity**

813 "*A scientific article that presents all of the data and all of the scientific discourse that the author*
814 *intended to present is not necessarily a successful article. It only becomes one when most of the*
815 *people who read it can perceive accurately and quickly what the author really meant*." (Lindsay
816 2011, p. 64). Reading papers that are not written clearly and precisely is laborious and time
817 consuming, and when pressed for time, we often end up not being able to read these papers in
818 detail. This can result in papers that present important and interesting research not receiving the
819 attention and credit they deserve.

820

821 Writing clearly and precisely is not necessarily a question of language proficiency. In his book
822 *Scientific Writing = Thinking in Words*, David Lindsay (2011, p. 12) writes:

823

824 "*Science has a language of its own that has nothing to do with the scientist's native tongue. It is*
825 *the language of logic in which reasoned arguments are developed from well-presented evidence*
826 *and lead to sound and consistent conclusions. That language is the same regardless of the origin*
827 *and preferred tongue of the person who writes it and good scientific writing depends primarily*
828 *on expressing the science precisely and clearly. Subsequent editing by a native speaker to tidy up*
829 *English expressions and comply with modern vernacular is relatively easy and the article will be*
830 *a good one*."

831

832 Numerous authors have written in detail on how to produce high-quality scientific writing (notably
833 Alley 1996; O'Connor 2002; Gustavii 2016). Here we summarize advice from Lindsay (2011) on

how to build a logical structure for a scientific paper, how to achieve clarity and flow by presenting the relevant information in the right order, and how to make simple but powerful stylistic adjustments to enable the reader to grasp the meaning quickly and accurately. It is beyond the scope of this paper to summarize all of the key advice. In a plea for scientific practitioners—both those who are comfortable with writing, and those who find writing challenging—to consult this and other resources, we point out three main pieces of advice.

(a) The Introduction is the "powerhouse of your article" (p.20). It orientates the reader within the wider discussion to which the authors wish to contribute new knowledge. It does not just provide the theoretical or methodological background to the study. It provides the foundation of the logical structure of the overarching argument and sets up the flow with which the Results and Discussion can be delivered effectively. To achieve these goals, writing the Introduction entails following two simple principles (p.21):

1. The hypothesis is key to the Introduction.
2. By justifying the hypothesis logically and scientifically, the writer provides just about everything necessary for readers to understand what the paper is about and why the author wrote it.

Although many writers prefer to write the Introduction last, given its central role in the logical structure of the argument, Lindsay (2011) convincingly demonstrates why it can be beneficial to start writing it first.

(b) The Introduction, Results, and Discussion work together. For the Discussion to be meaningful and convincing, the reader must be prepared (from the Introduction) to understand how the evidence gathered in this study (Results) logically feeds into the argument. To achieve their impact, results should be presented in the order of importance. Thus, rather than saving the most important for the end to create suspense, the most important finding should be presented first. To achieve this, Lindsay (2011) suggests that scientists sort their results into four categories (p.31):

1. Results that are clear and relevant to what you want to say about the hypothesis.
2. Results that allow you to say something relevant about the hypothesis but that are less convincing than the results in Category 1.

3. Results that are interesting, substantial, and worth presenting, but they don't have anything to do with the hypothesis.

4. Results that are not convincing and don't have anything to do with the hypothesis.

Once sorted, the results can be easily transformed into scaffolding for the logical argument. Category 4 results should be omitted, while Category 1–3 should be presented in exactly this order.

(c) Following the mantra that scientific writing should 'inform, not impress' will automatically result in avoiding some of the stumbling blocks that frequently characterize scientific writing" (p.55). This will solve two problems. Firstly, it will prevent the Discussion from becoming too long as a result of inclusion of unnecessary references and deliberations that do not lead to the conclusions being made (which, according to editors of scientific journals, are some of the most common problems of scientific manuscripts). Secondly, it will free the language from problems of clarity and brevity. Lindsay (2011) explains and provides examples for how to avoid seven main types of writing stumbling blocks: clusters of nouns; complex adjectival phrases; sentences beginning with subordinate clauses; using nouns instead of the verbs from which they are derived; use of imprecise words; use of acronyms; unfamiliar abbreviations, and symbols; footnotes, asides in parentheses, and other distractions. Examples of these stumbling blocks from isotope archaeology are discussed in Supplementary Material 3. In addition, the author lays out five steps which can be used to check the clarity of every paragraph: Is it a paragraph?; Do the sentences flow?; Are there stumbling blocks?; Can it be shortened without losing the meaning?; Does it say what you want it to say?. Due to space constraints, we are unable to provide examples for all these points of advice, but reiterate our conviction that both seasoned and up-and-coming writers will benefit from engaging with Lindsay's short book.

6. Conclusions

This paper has presented suggestions and opportunities for improvement in three areas of project design and implementation: selecting samples, processing data and disseminating results. Some of the suggestions are relatively easy to implement:

- Consider carrying out power analyses to determine optimal sample sizes
- Invest time into exploring the distribution of data prior to formal statistical analysis

- Consider issues like overcrowding and colorblind-friendly palettes when preparing graphical figures
- Do not use the term ‘statistically significant difference’ (or any paraphrased equivalents)
- Do not hinge interpretation of group differences on *p*-value thresholds
- Consider measurement uncertainty and sample type when deciding how many decimal places to report; and report this number consistently throughout a publication
- Follow the established conventions and terminology for reporting stable isotope values ($\delta^{13}C$, $\delta^{15}N$, $\delta^{18}O$) and isotopic ratios ($^{87}Sr/^{86}Sr$)
- Take steps to justify sample selection, preservation and integrity (e.g., when reporting analyses of plants, include a supplementary document with photographs showing the surface morphology and internal structure of all analyzed seeds; when reporting the analyses of shells, bones and teeth, provide a table with results of all chemical analyses undertaken during sample screening)
- Consider how $\delta^{13}C$ values are obtained during AMS radiocarbon dating when deciding whether to use the values for answering palaeodietary/palaeoclimatic questions

Some will require an initial investment in training, which will pay off in the future:

- Assess measurement error by calculating both precision and accuracy (using measured values of both calibration and check standards inserted into each analytical run) and provide all relevant details in supplementary materials to enable the reader to replicate the calculations independently
- Undertake targeted analyses of shell carbonate structure to determine their integrity prior to analysis
- Assess both the surface morphology and internal structure (shown in cross-sections) of charred seeds to determine their likelihood of preserving usable isotopic values

Some will require an ongoing investment in time and effort:

- Consult the evolving literature on the latest advice for quality control assessment
- Consult with professional statisticians on the design of experiments and collection of data
- Before analyzing crop species that have not been analyzed before, carry out charring experiments to determine the crops’ preservation potential

Although isotopic research in archaeology requires a group effort and not everyone involved needs to understand the intricacies of the method, time investment from all parties involved can make a difference on both the final numbers and their interpretations. When researchers who cannot run analyses on their own are given full and transparent descriptions of the laboratory protocols and workflows from the collaborating laboratories and technicians, they will be better informed about where possible sources of error can occur. When archaeologists who have not been trained in isotopic geochemistry, but who are interested in collaborating on isotopic analyses, try their hand at data calibration and other data handling, they will be better positioned to communicate the reliability of the results. When students make time to practice (or at least engage in discussions surrounding) conventions and best practices such as those suggested in this paper, they will become effective in communicating the results both verbally and in writing.

Close collaboration with professional statisticians is crucial for avoiding mistakes that have been, and are continuing to be made, by the wider scientific community. In addition, we encourage isotope practitioners to write rebuttal papers more frequently, to participate in sharing data on Open Access repositories (e.g., IsoArch, IsoMemo), and to interact with papers during the revision stage (by reading and commenting on 'pre-prints' online). Lastly, many people who speak to, collaborate with, or receive feedback from Julia Lee-Thorp have heard her say "Why would you want to do that?" in a variety of contexts. To follow in the footsteps of her healthy dose of skepticism, we urge researchers to keep this question in mind during all stages of research. At no point should there be a time when they cannot answer it, or find out where the answer would be.

Author contributions

Conceptualization: PV, NL; Data curation: N/A; Formal analysis: PV, NL; Funding acquisition: N/A; Investigation: N/A; Methodology: N/A; Project administration: N/A; Resources: N/A; Software: N/A; Supervision: N/A; Validation: N/A; Visualization: PV, NL; Roles/Writing - original draft: PV, NL, ES, EL, CM; Writing - review & editing: PV, NL, ES, EL, CM.

Data availability

No original data was used in this paper. Stable isotope data that was re-examined in Section 5 was drawn from Vaiglova et al. (2020) and can be found at: https://link.springer.com/article/10.1007/s12520-019-00960-y#additional-information.

Declaration of competing interest

The authors declare no competing interests.

Acknowledgements

We are thankful to Richard Staff, David Chivall, Peter Ditchfield, Patrick Roberts, Maura Pellegrini, Melanie Beasley, Corrie Hyland, Rick Schulting, and Melissa Ritchey for their helpful comments and suggestions.

References

Aldeias V, Gur-Arieh S, Maria R, et al. (2019) Shell we cook it? An experimental approach to the microarchaeological record of shellfish roasting. Archaeol Anthropol Sci 11:389–407.

Alley M (1996) The craft of scientific writing. Springer, New York.

Alves EQ, Macario KD, Urrutia FP, et al. (2019) Accounting for the marine reservoir effect in radiocarbon calibration. Quat Sci Rev 209:129–138.

Ambrose SH (1990) Preparation and characterization of bone and tooth collagen for isotopic analysis. J Archaeol Sci 17:431–451.

Ambrose SH, Norr L (1993) Experimental Evidence for the Relationship of the Carbon Isotope Ratios of Whole Diet and Dietary Protein to Those of Bone Collagen and Carbonate. In: Lambert JB, Grupe G (eds) Prehistoric Human Bone: Archaeology at the Molecular Level. Springer, Berlin, pp 1–37.

Andrus CFT (2011) Shell midden sclerochronology. Quat Sci Rev 30:2892–2905.

Atalay S (2010) 'We don't talk about Çatalhöyük, we live it': sustainable archaeological practice through community-based participatory research. World Archaeol 42:418–429.

Aubert M, Williams IS, Boljkovac K, et al. (2012) In situ oxygen isotope micro-analysis of faunal material and human teeth using a SHRIMP II: a new tool for palaeo-ecology and archaeology. J Archaeol Sci 39:3184–3194.

Austin C, Smith TM, Farahani RMZ, et al. (2016) Uncovering system-specific stress signatures in primate teeth with multimodal imaging. Sci Rep 6:18802.

Balasse M (2003) Potential biases in sampling design and interpretation of intra-tooth isotope analysis. Int J Osteoarchaeol 13:3–10.

Balasse M, Bălăşescu A, Janzen A, et al. (2013) Early herding at Măgura-Boldul lui Moş Ivănuş (early sixth millennium BC, Romania): environments and seasonality from stable isotope analysis. Eur J Archaeol 16:221–246.

Balasse M, Bocherens H, Mariotti A, Ambrose SH (2001) Detection of Dietary Changes by Intra-tooth Carbon and Nitrogen Isotopic Analysis: An Experimental Study of Dentine Collagen of Cattle (Bos taurus). J Archaeol Sci 28:235–245.

Balasse M, Boury L, Ughetto-Monfrin J, Tresset A (2012) Stable isotope insights ($\delta^{18}O$, $\delta^{13}C$) into cattle and sheep husbandry at Bercy (Paris, France, 4th millennium BC): birth seasonality

and winter leaf foddering. Environ Archaeol 17:29–44.
Balasse M, Gillis R, Živaljević I, et al. (2021) Seasonal calving in European Prehistoric cattle and its impacts on milk availability and cheese-making. Sci Rep 11:1–11.
Beasley MM, Bartelink EJ, Taylor L, Miller RM (2014) Comparison of transmission FTIR, ATR, and DRIFT spectra: implications for assessment of bone bioapatite diagenesis. J Archaeol Sci 46:16–22.
Birch J (2012) Worldwide prevalence of red-green color deficiency. JOSA A 29:313–320.
Bogaard A, Heaton THE, Poulton P, Merbach I (2007) The impact of manuring on nitrogen isotope ratios in cereals: archaeological implications for reconstruction of diet and crop management practices. J Archaeol Sci 34:335–343.
Bonafini M, Pellegrini M, Ditchfield P, Pollard AM (2013) Investigation of the 'canopy effect' in the isotope ecology of temperate woodlands. J Archaeol Sci 40:3926–3935.
Bond AL, Hobson KA (2012) Reporting Stable-Isotope Ratios in Ecology: Recommended Terminology, Guidelines and Best Practices. Waterbirds 35:324–331.
Brand U (1989) Aragonite-calcite transformation based on Pennsylvanian molluscs. Geol Soc of Am Bull 101:377–390.
Cavazzuti C, Skeates R, Millard AR, et al. (2019) Flows of people in villages and large centres in Bronze Age Italy through strontium and oxygen isotopes. Plos One 14:e0209693.
Cernusak LA, Ubierna N, Winter K, et al. (2013) Environmental and physiological determinants of carbon isotope discrimination in terrestrial plants. New Phytol 200:950–965.
Charles M, Forster E, Wallace M, Jones G (2015) "Nor ever lightning char thy grain": establishing archaeologically relevant charring conditions and their effect on glume wheat grain morphology. STAR Sci Technol Archaeol Res 1:1–6.
Chesson LA, Beasley MM, Bartelink EJ, et al. (2021) Using bone bioapatite yield for quality control in stable isotope analysis applications. J Archaeol Sci Rep 35:102749.
Cheung C, Szpak P (2021) Interpreting Past Human Diets Using Stable Isotope Mixing Models—Best Practices for Data Acquisition. J Archaeol Method Theory 1–24.
Clementz MT (2012) New insight from old bones: stable isotope analysis of fossil mammals. J Mammal 93:368–380.
Cleveland WS (1994) The Elements of Graphing Data. Hobart Press, New Jersey.
Codron D, Hull J, Brink JS, et al (2011) Effect of competition on niche dynamics of syntopic grazing ungulates: contrasting the predictions of habitat selection models using stable isotope analysis. Evol Ecol Res 13:217–235.
Codron J, Lee-Thorp JA, Sponheimer M, Codron D (2013) Plant stable isotope composition across habitat gradients in a semi-arid savanna: implications for environmental reconstruction. J Quat Sci 28:301–310.
Collard M, Wasef S, Adams S, et al. (2019) Giving it a burl: towards the integration of genetics, isotope chemistry, and osteoarchaeology in Cape York, Tropical North Queensland, Australia. World Archaeol 51:602–619.
Copeland SR, Sponheimer M, le Roux PJ, et al. (2008) Strontium isotope ratios ($^{87}Sr/^{86}Sr$) of tooth enamel: a comparison of solution and laser ablation multicollector inductively coupled plasma mass spectrometry methods. Rapid Commun Mass Spectrom 22:3187–3194.
Coplen TB (2011) Guidelines and recommended terms for expression of stable-isotope-ratio and gas-ratio measurement results. Rapid Commun Mass Spectrom 25:2538–560.
Cucchi T, Dai L, Balasse M, et al. (2016) Social Complexification and Pig (*Sus scrofa*) Husbandry in Ancient China: A Combined Geometric Morphometric and Isotopic Approach.

Plos One 11:e0158523.
DeNiro MJ (1985) Postmortem preservation and alteration of in vivo bone collagen isotope ratios in relation to palaeodietary reconstruction. Nature 317:806–809.
DeNiro MJ, Weiner S (1988) Chemical, enzymatic and spectroscopic characterization of "collagen" and other organic fractions from prehistoric bones. Geochim Cosmochim Acta 52:2197–2206.
Dobberstein RC, Collins MJ, Craig OE, et al. (2009) Archaeological collagen: Why worry about collagen diagenesis? Archaeol Anthropol Sci 1:31–42.
Ehrenberg ASC (1981) The Problem of Numeracy. Am Stat 35:67–71.
Ericson JE (1985) Strontium isotope characterization in the study of prehistoric human ecology. J Hum Evol 14:503–514.
Evans JA, Montgomery J, Wildman G, Boulton N (2010) Spatial variations in biosphere $^{87}Sr/^{86}Sr$ in Britain. J Geol Soc 167:1–4.
Fernandes R (2016) A Simple(R) Model to Predict the Source of Dietary Carbon in Individual Consumers. Archaeometry 58:500–512.
Flohr P, Jenkins E, Williams HRS, et al. (2019) What can crop stable isotopes ever do for us? An experimental perspective on using cereal carbon stable isotope values for reconstructing water availability in semi-arid and arid environments. Veg Hist Archaeobotany 28:497–512.
France CAM, Sugiyama N, Aguayo E (2020) Establishing a preservation index for bone, dentin, and enamel bioapatite mineral using ATR-FTIR. J Archaeol Sci Rep 33:102551.
Fraser RA, Bogaard A, Heaton T, et al. (2011) Manuring and stable nitrogen isotope ratios in cereals and pulses: towards a new archaeobotanical approach to the inference of land use and dietary practices. J Archaeol Sci 38:2790–2804.
Frémondeau D, Cucchi T, Casabianca F, et al. (2012) Seasonality of birth and diet of pigs from stable isotope analyses of tooth enamel ($\delta^{18}O$, $\delta^{13}C$): a modern reference data set from Corsica, France. J Archaeol Sci 39:2023–2035.
Fry B (2006) Stable Isotope Ecology. Springer, Baton Rouge.
Galimberti M, Loftus E, Sealy J (2017) Investigating $\delta^{18}O$ of *Turbo sarmaticus* (L. 1758) as an indicator of sea surface temperatures. Palaeogeogr Palaeoclimatol Palaeoecol 484:62–69.
Gannes LZ, O'Brien DM, del Rio CM (1997) Stable Isotopes in Animal Ecology: Assumptions, Caveats, and a Call for More Laboratory Experiments. Ecology 78:1271–1276.
Gillis RE, Bulatović J, Penezić K, et al. (2021) Of herds and societies—Seasonal aspects of Vinča culture herding and land use practices revealed using sequential stable isotope analysis of animal teeth. Plos One 16:e0258230.
Gillis RE, Eckelmann R, Filipović D, et al. (2020) Stable isotopic insights into crop cultivation, animal husbandry, and land use at the Linearbandkeramik site of Vráble-Veľké Lehemby (Slovakia). Archaeol Anthropol Sci 12:256.
Goodfriend GA (1992) The use of land snail shells in paleoenvironmental reconstruction. Quat Sci Rev 11:665–685.
Goodman S (2008) A Dirty Dozen: Twelve P-Value Misconceptions. Semin Hematol 45:135–140.
Greenland S, Senn SJ, Rothman KJ, et al. (2013) Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 31:337–350.
Guiry EJ, Szpak P (2021) Improved quality control criteria for stable carbon and nitrogen isotope measurements of ancient bone collagen. J Archaeol Sci 132:105416.
Guiry EJ, Szpak P, Richards MP (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.
Gustavii B (2016) How to Write and Illustrate a Scientific Paper, 2nd edn. Cambridge University Press, Cambridge.
Hart JP, Feranec RS (2020) Using Maize δ^{15}N values to assess soil fertility in fifteenth- and sixteenth-century ad Iroquoian agricultural fields. Plos One 15:e0230952.
Head ML, Holman L, Lanfear R, et al. (2015) The Extent and Consequences of P-Hacking in Science. Plos Biol 13:e1002106.
Hillson S (2005) Teeth. Cambridge university press, Cambridge.
Holt E, Evans JA, Madgwick R (2021) Strontium (^{87}Sr/^{86}Sr) mapping: A critical review of methods and approaches. Earth-Sci Rev 216:103593.
Hoppe KA, Stover SM, Pascoe JR, Amundson R (2004) Tooth enamel biomineralization in extant horses: implications for isotopic microsampling. Palaeogeogr Palaeoclimatol Palaeoecol 206:355–365.
Hyland C, Scott MB, Routledge J, Szpak P (2021) Stable Carbon and Nitrogen Isotope Variability of Bone Collagen to Determine the Number of Isotopically Distinct Specimens. J Archaeol Method Theory.
Ivany LC (2012) Reconstructing Paleoseasonality from Accretionary Skeletal Carbonates—Challenges and Opportunities. Paleontol Soc Pap 18:133–166.
Janzen A, Balasse M, Ambrose SH (2020) Early pastoral mobility and seasonality in Kenya assessed through stable isotope analysis. J Archaeol Sci 117:1–17.
Johnson BJ, Fogel ML, Miller GH (1998) Stable isotopes in modern ostrich eggshell: a calibration for paleoenvironmental applications in semi-arid regions of southern Africa. Geochim Cosmochim Acta 62:2451–2461.
Joint Committee for Guides in Metrology (2008) Evaluation of Measurement Data – Guide to the Expression of Uncertainty in Measurement. International Organization for Standardization, Geneva.
Jones AM, O'Connell TC, Young ED, et al. (2001) Biogeochemical data from well preserved 200 ka collagen and skeletal remains. Earth Planet Sci Lett 193:143–149.
Jones PJ, O'Connell TC, Jones MK, et al (2021) Crop water status from plant stable carbon isotope values: A test case for monsoonal climates. The Holocene 31:993–1004.
Knipper C, Pichler SL, Rissanen H, et al. (2017) What is on the menu in a Celtic town? Iron Age diet reconstructed at Basel-Gasfabrik, Switzerland. Archaeol Anthropol Sci 9:1307–1326.
Knipper C, Reinhold S, Gresky J, et al. (2020) Diet and subsistence in Bronze Age pastoral communities from the southern Russian steppes and the North Caucasus. Plos one 15:e0239861.
Knudson KJ, Stanish C, Lozada Cerna MC, et al. (2016) Intra-individual variability and strontium isotope measurements: A methodological study using ^{87}Sr/^{86}Sr data from Pampa de los Gentiles, Chincha Valley, Peru. J Archaeol Sci Rep 5:590–597.
Koch PL, Tuross N, Fogel ML (1997) The Effects of Sample Treatment and Diagenesis on the Isotopic Integrity of Carbonate in Biogenic Hydroxylapatite. J Archaeol Sci 24:417–429.
Lee-Thorp J, Sponheimer M (2003) Three case studies used to reassess the reliability of fossil bone and enamel isotope signals for paleodietary studies. J Anthropol Archaeol 22:208–216.
Lee-Thorp J, Van der Merwe NJ (1987) Carbon isotope analysis of fossil bone apatite. South Afr J Sci 83:712–715.
Lee-Thorp JA (2008) On isotopes and old bones. Archaemetry 50:925–950.
Leng MJ, Lewis JP (2014) Oxygen isotopes in Molluscan shell: Applications in environmental

archaeology. Environ Archaeol 21:295–306.
Levin NE, Brown FH, Behrensmeyer AK, et al. (2011) Paleosol carbonates from the Omo Group: Isotopic records of local and regional environmental change in East Africa. Palaeogeogr Palaeoclimatol Palaeoecol 307:75–89.
Lindsay D (2011) Scientific Writing = Thinking in Words. Csiro Publishing, Collingwood.
Loftus E, Rogers K, Lee-Thorp J (2015) A simple method to establish calcite:aragonite ratios in archaeological mollusc shells. J Quat Sci 30:731–735.
Loftus E, Rogers K, Lee-Thorp J (2015) A simple method to establish calcite:aragonite ratios in archaeological mollusc shells. J Quat Sci 30:731–735.
Lowenstam HA (1954) Factors Affecting the Aragonite: Calcite Ratios in Carbonate-Secreting Marine Organisms. J Geol 62:284–322.
Macfadden BJ, Cerling TE (1996) Mammalian herbivore communities, ancient feeding ecology, and carbon isotopes: A 10 million-year sequence from the Neogene of Florida. J Vertebr Paleontol 16:103–115.
Makarewicz CA, Pederzani S (2017) Oxygen (δ^{18}O) and carbon (δ^{13}C) isotopic distinction in sequentially sampled tooth enamel of co-localized wild and domesticated caprines: Complications to establishing seasonality and mobility in herbivores. Palaeogeogr Palaeoclimatol Palaeoecol 485:1–15.
Marston JM (2021) Archaeological Approaches to Agricultural Economies. J Archaeol Res 29:327–385.
Masters PM (1987) Preferential preservation of noncollagenous protein during bone diagenesis: Implications for chronometric and stable isotopic measurements. Geochim Cosmochim Acta 51:3209–3214.
Milano S, Prendergast AL, Schöne BR (2016) Effects of cooking on mollusk shell structure and chemistry: Implications for archeology and paleoenvironmental reconstruction. J Archaeol Sci Rep 7:14–26.
Miller RG (1986) Beyond ANOVA: Basics of Applied Statistics. John Wiley and Sons, New York.
Müldner G, Richards MP (2007) Stable isotope evidence for 1500 years of human diet at the city of York, UK. Am J Phys Anthropol 133:682–697.
Murphy KR, Myors B, Wolach A (2014) Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests, Fourth Edition, 4th edn. Routledge, New York.
Nitsch E, Andreou S, Creuzieux A, et al. (2017) A bottom-up view of food surplus: using stable carbon and nitrogen isotope analysis to investigate agricultural strategies and diet at Bronze Age Archontiko and Thessaloniki Toumba, northern Greece. World Archaeol 49:105–137.
Nitsch EK, Charles M, Bogaard A (2015) Calculating a statistically robust δ^{13}C and δ^{15}N offset for charred cereal and pulse seeds. Sci Technol Archaeol Res 1:1–8.
O'Connor M (2002) Writing Successfully in Science. Routledge, London.
Pálsdóttir AH, Bläuer A, Rannamäe E, et al. (2019) Not a limitless resource: ethics and guidelines for destructive sampling of archaeofaunal remains. R Soc Open Sci 6:191059.
Parnell AC, Phillips DL, Bearhop S, et al. (2013) Bayesian stable isotope mixing models. Environmetrics 24:387–399.
Pearson J, Grove M (2013) Counting sheep: sample size and statistical inference in stable isotope analysis and palaeodietary reconstruction. World Archaeol 45:373–387.
Pechenkina EA, Ambrose SH, Xiaolin M, Benfer RA (2005) Reconstructing northern Chinese

Neolithic subsistence practices by isotopic analysis. J Archaeol Sci 32:1176–1189.
Pederzani S, Britton K (2019) Oxygen isotopes in bioarchaeology: Principles and applications, challenges and opportunities. Earth-Sci Rev 188:77–107.
Pollard AM, Pellegrini M, Lee-Thorp JA (2011) Technical note: Some observations on the conversion of dental enamel $\delta^{18}O_p$ values to $\delta^{18}O_w$ to determine human mobility. Am J Phys Anthropol 145:499–504.
Price M, Rowan YM, Kersel MM, Makarewicz CA (2020) Fodder, pasture, and the development of complex society in the Chalcolithic: isotopic perspectives on animal husbandry at Marj Rabba. Archaeol Anthropol Sci 12:95.
Pryor AJE, Insoll T, Evis L (2020) Laser ablation strontium isotope analysis of human remains from Harlaa and Sofi, eastern Ethiopia, and the implications for Islamisation and mobility. STAR Sci Technol Archaeol Res 6:113–136.
Reimer PJ, Brown TA, Reimer RW (2004) Discussion: Reporting and Calibration of Post-Bomb 14C Data. Radiocarbon 46:1299–1304.
Reyes-Gasga J, Martínez-Piñeiro EL, Rodríguez-Álvarez G, et al. (2013) XRD and FTIR crystallinity indices in sound human tooth enamel and synthetic hydroxyapatite. Mater Sci Eng C 33:4568–4574.
Roberts P, Fernandes R, Craig OE, et al. (2017a) Calling all archaeologists: guidelines for terminology, methodology, data handling, and reporting when undertaking and reviewing stable isotope applications in archaeology. Rapid Commun Mass Spectrom 32:361–372.
Roberts P, Perera N, Wedage O, et al. (2017b) Fruits of the forest: Human stable isotope ecology and rainforest adaptations in Late Pleistocene and Holocene (~36 to 3 ka) Sri Lanka. J Hum Evol 106:102–118.
Roberts P, Gillingwater TH, Lahr MM, et al (2018) Historical Tropical Forest Reliance amongst the Wanniyalaeto (Vedda) of Sri Lanka: an Isotopic Perspective. Hum Ecol 46:435–444.
Roberts P, Perera N, Wedage O, et al. (2015) Direct evidence for human reliance on rainforest resources in late Pleistocene Sri Lanka. Science 347:1246–1249.
Roffet-Salque M, Marciniak A, Valdes PJ, et al. (2018) Evidence for the impact of the 8.2-kyBP climate event on Near Eastern early farmers. Proc Natl Acad Sci 115:8705–8709.
Rokita E, Hermes C, Nolting H-F, Ryczek J (1993) Substitution of calcium by strontium within selected calcium phosphates. J Cryst Growth 130:543–552.
Royer A, Daux V, Fourel F, Lécuyer C (2017) Carbon, nitrogen and oxygen isotope fractionation during food cooking: Implications for the interpretation of the fossil human record. Am J Phys Anthropol 163:759–771.
Rull V, Vegas-Vilarrúbia T, Corella JP, Valero-Garcés B (2021) Bronze Age to Medieval vegetation dynamics and landscape anthropization in the southern-central Pyrenees. Palaeogeogr Palaeoclimatol Palaeoecol 571:110392.
Schoeninger MJ, Moore KM, Murray ML, Kingston JD (1989) Detection of bone preservation in archaeological and fossil samples. Appl Geochem 4:281–292.
Schöne BR (2008) The curse of physiology—challenges and opportunities in the interpretation of geochemical data from mollusk shells. Geo-Mar Lett 28:269–285.
Sharp Z (2017) Principles of Stable Isotope Geochemistry. Pearson, online only edition.
Shemesh A (1990) Crystallinity and diagenesis of sedimentary apatites. Geochim Cosmochim Acta 54:2433–2438.
Simmons JP, Nelson LD, Simonsohn U (2011) False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. Psychol

Sci 22:1359–1366.
Smith TM, Austin C, Green DR, et al (2018) Wintertime stress, nursing, and lead exposure in Neanderthal children. Sci Adv 4:1–9.
Smith CI, Nielsen-Marsh CM, Jans MME, Collins MJ (2007) Bone diagenesis in the European Holocene I: patterns and mechanisms. J Archaeol Sci 34:1485–1493.
Snoeck C, Lee-Thorp J, Schulting R, et al. (2015) Calcined bone provides a reliable substrate for strontium isotope ratios as shown by an enrichment experiment. Rapid Commun Mass Spectrom 29:107–114.
Snoeck C, Ryan S, Pouncett J, et al. (2020) Towards a biologically available strontium isotope baseline for Ireland. Sci Total Environ 712:136248.
Snoeck C, Schulting RJ, Lee-Thorp JA, et al. (2016) Impact of heating conditions on the carbon and oxygen isotope composition of calcined bone. J Archaeol Sci 65:32–43.
Sponheimer M, Passey BH, de Ruiter DJ, et al. (2006) Isotopic Evidence for Dietary Variability in the Early Hominin Paranthropus robustus. Science 314:980–982.
Squires K, Booth T, Roberts CA (2019) The Ethics of Sampling Human Skeletal Remains for Destructive Analyses. Ethical Approaches Hum Remains 265–297.
Stevens RE, Lister AM, Hedges REM (2006) Predicting diet, trophic level and palaeoecology from bone stable isotope analysis: a comparative study of five red deer populations. Oecologia 149:12–21.
Stuiver M, Polach HA (1977) Discussion Reporting of ^{14}C Data. Radiocarbon 19:355–363.
Styring A, Maier U, Stephan E, et al. (2016) Cultivation of choice: new insights into farming practices at Neolithic lakeshore sites. Antiquity 90:95–110.
Styring AK, Charles M, Fantone F, et al. (2017) Isotope evidence for agricultural extensification reveals how the world's first cities were fed. Nat Plants 3:17076.
Styring AK, Evans JA, Nitsch EK, et al. (2019) Revisiting the potential of carbonized grain to preserve biogenic $^{87}Sr/^{86}Sr$ signatures within the burial environment. Archaeometry 61:179–193.
Sullivan CH, Krueger HW (1981) Carbon isotope analysis of separate chemical phases in modern and fossil bone. Nature 292:333–335.
Syväranta J, Lensu A, Marjomäki TJ, et al. (2013) An Empirical Evaluation of the Utility of Convex Hull and Standard Ellipse Areas for Assessing Population Niche Widths from Stable Isotope Data. Plos One 8:e56094.
Szpak P, Chiou KL (2019) A comparison of nitrogen isotope compositions of charred and desiccated botanical remains from northern Peru. Veg Hist Archaeobotany 29:527–538.
Szpak P, Metcalfe JZ, Macdonald RA (2017) Best practices for calibrating and reporting stable isotope measurements in archaeology. J Archaeol Sci Rep 13:609–616.
Thomson JJ, Aston FW, Soddy F, et al. (1921) Discussion on Isotopes. Proc R Soc Lond Ser Contain Pap Math Phys Character 99:87–104.
Tieszen LL (1991) Natural variations in the carbon isotope values of plants: implications for archaeology, ecology, and paleoecology. J Archaeol Sci 18:227–248.
Vaiglova P, Gardeisen A, Buckley M, et al. (2020a) Further insight into Neolithic agricultural management at Kouphovouno, southern Greece: expanding the isotopic approach. Archaeol Anthropol Sci 12:43.
Vaiglova P, Hartman G, Marom N, et al. (2020b) Climate stability and societal decline on the margins of the Byzantine empire in the Negev Desert. Sci Rep 10:1512.
Vaiglova P, Halstead P, Pappa M, et al. (2018) Of cattle and feasts: Multi-isotope investigation of animal husbandry and communal feasting at Neolithic Makriyalos, northern Greece. Plos One

13:e0194474.
Vaiglova P, Reid REB, Lightfoot E, et al. (2021) Localized management of non-indigenous animal domesticates in Northwestern China during the Bronze Age. Sci Rep 11:15764.
Vaiglova P, Snoeck C, Nitsch E, et al. (2014) Impact of contamination and pre- treatment on stable carbon and nitrogen isotopic composition of charred plant remains. Rapid Commun Mass Spectrom 28:2497–2510.
van Klinken GJ (1999) Bone Collagen Quality Indicators for Palaeodietary and Radiocarbon Measurements. J Archaeol Sci 26:687–695.
Vogel JC, van der Merwe NJ (1977) Isotopic Evidence for Early Maize Cultivation in New York State. Am Antiq 42:238–242.
Wainer H (2007) Graphic Discovery: A Trout in the Milk and Other Visual Adventures. Princeton University Press, Princeton.
Wallace M, Jones G, Charles M, et al. (2013) Stable carbon isotope analysis as a direct means of inferring crop water status and water management practices. World Archaeol 45:388–409.
Wang X, Roberts P, Tang Z, et al. (2021) The Circulation of Ancient Animal Resources Across the Yellow River Basin: A Preliminary Bayesian Re-evaluation of Sr Isotope Data From the Early Neolithic to the Western Zhou Dynasty. Front Ecol Evol 9:16.
Wasserstein RL, Lazar NA (2016) The ASA Statement on p-values: Context, Process, and Purpose. Am Stat 70:129–133.
Wasserstein RL, Schirm AL, Lazar NA (2019) Moving to a World Beyond "p < 0.05." Am Stat 73:1–19.
Webb EC, White CD, Longstaffe FJ (2014) Investigating inherent differences in isotopic composition between human bone and enamel bioapatite: implications for reconstructing residential histories. J Archaeol Sci 50:97–107.
Weber K, Weber M, Menneken M, et al. (2021) Diagenetic stability of non-traditional stable isotope systems (Ca, Sr, Mg, Zn) in teeth – An in-vitro alteration experiment of biogenic apatite in isotopically enriched tracer solution. Chem Geol 572:120196.
Weiner S, Bar-Yosef O (1990) States of preservation of bones from prehistoric sites in the Near East: A survey. J Archaeol Sci 17:187–196.
Weinreb MM (1964) Tooth development in sheep. Am J Vet Res 25:891–908.
Weissgerber TL, Winham SJ, Heinzen EP, et al. (2019) Reveal, Don't Conceal. Circulation 140:1506–1518.
Wickham H (2016) ggplot2: elegant graphics for data analysis 2016, 2nd edn. Springer.
Wierzbowski H, Anczkiewicz R, Pawlak J, et al. (2017) Revised Middle–Upper Jurassic strontium isotope stratigraphy. Chem Geol 466:239–255.
Wilkinson L (2005) The Grammar of Graphics. In: Handbook of Computational Statistics: Concepts and Methods, 2nd edn. Springer, New York.
Wood R, Fleury ABC, Fallon S, et al. (2021) Do weak or strong acid remove carbonate contamination from ancient tooth enamel more effectively? The effect of acid pretreatment on radiocarbon and $\delta^{13}C$ analyses. Radiocarbon 63:935–952.
Wright LE, Schwarcz HP (1996) Infrared and Isotopic Evidence for Diagenesis of Bone Apatite at Dos Pilas, Guatemala: Palaeodietary Implications. J Archaeol Sci 23:933–944.
Yanes Y, Izeta AD, Cattáneo R, et al. (2014) Holocene (~4.5–1.7 cal. kyr BP) paleoenvironmental conditions in central Argentina inferred from entire-shell and intra-shell stable isotope composition of terrestrial gastropods. The Holocene 24:1193–1205.
Yanes Y, Yapp CJ, Ibáñez M, et al. (2011) Pleistocene–Holocene environmental change in the

Canary Archipelago as inferred from the stable isotope composition of land snail shells. Quat Res 75:658–669.
Zazzo A, Balasse M, Patterson WP (2005) High-resolution $\delta^{13}C$ intratooth profiles in bovine enamel: Implications for mineralization pattern and isotopic attenuation. Geochim Cosmochim Acta 69:3631–3642.
Zazzo A, Lécuyer C, Sheppard SMF, et al. (2004) Diagenesis and the reconstruction of paleoenvironments: A method to restore original $\delta^{18}O$ values of carbonate and phosphate from fossil tooth enamel. Geochim Cosmochim Acta 68:2245–2258.
Zhang Q, Hou Y, Li X, et al (2021) Stable isotopes reveal intensive pig husbandry practices in the middle Yellow River region by the Yangshao period (7000–5000 BP). Plos One 16:e0257524.

Captions

Figure 1 Distribution of $\delta^{13}C$ values of the four main crops at Kouphovouno. Data was originally published in Vaiglova et al. (2020a) and the statistical analysis is being re-examined here.

Figure 2 Outcome of Tukey HSD comparison between the $\delta^{13}C$ values of the four main crops at Kouphovouno. Data was originally published in Vaiglova et al. (2020a) and the statistical analysis is being re-examined here.

Table 1 Summary of quality control criteria for archaeological materials most commonly analyzed in isotopic archaeology. Method development on these criteria is an active area of research, and the reader is encouraged to check the latest literature. References present experimental studies on quality control, unless preceded by "e.g.", in which case they present an example of an application of these criteria.

Table 2 Results of Tukey HSD comparisons between $\delta^{13}C$ values of free-threshing wheat, hulled barley, einkorn and pea. Data was originally published in Vaiglova et al. (2020a) and the statistical analysis is being re-examined here.

Figure 1

Click here to access/download;Figure;Fig.1.tiff

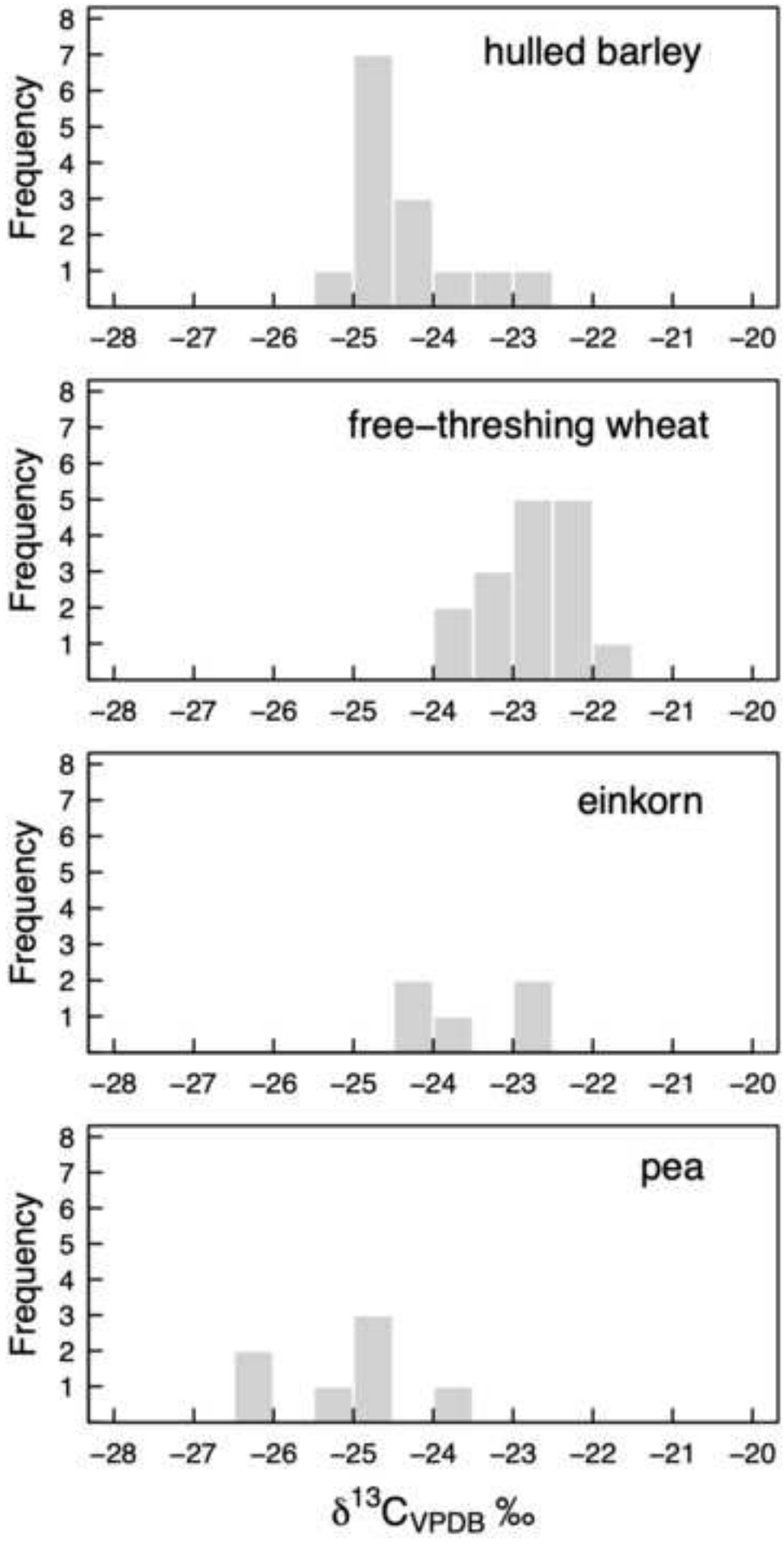
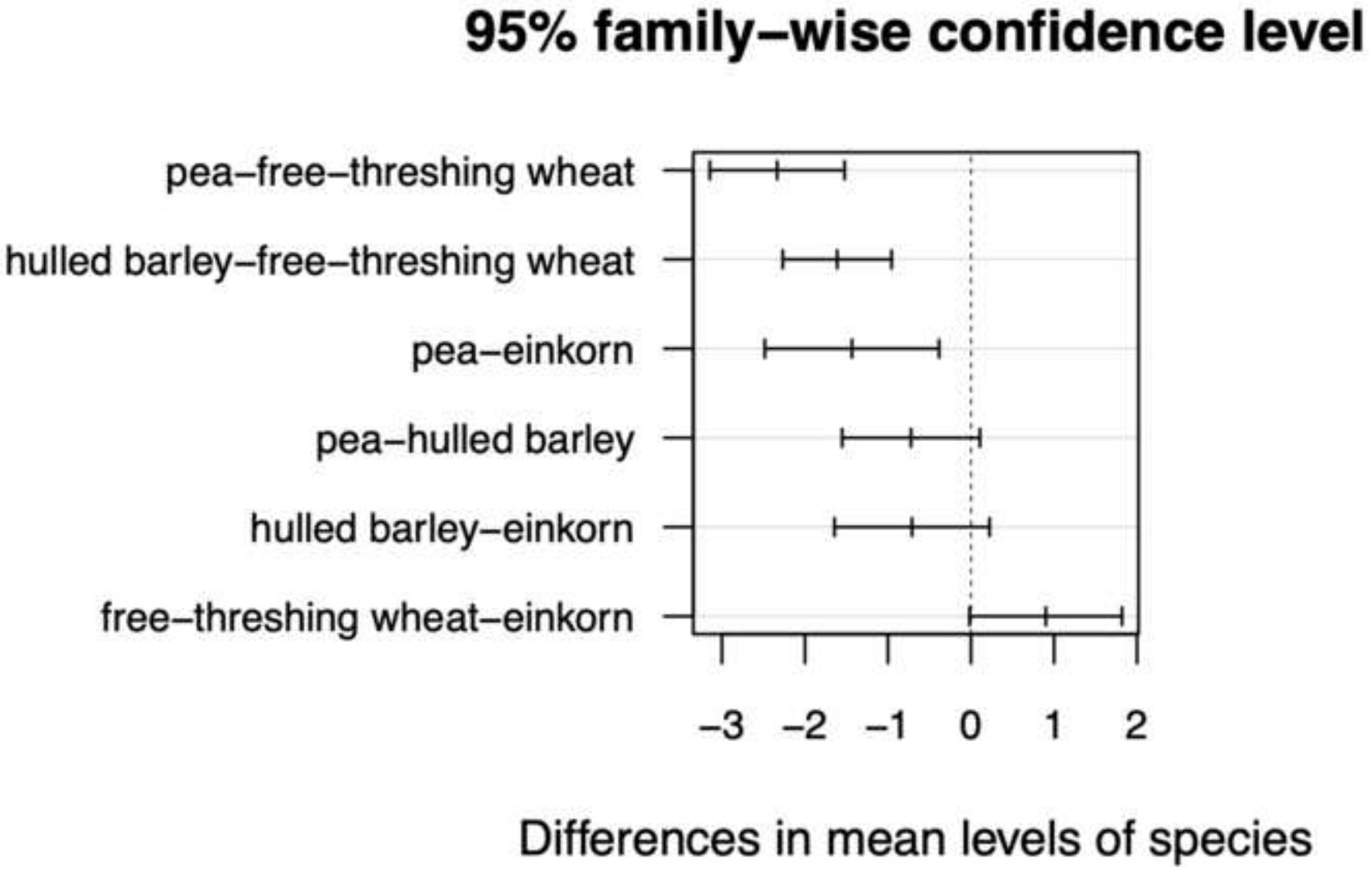


Figure 2

Click here to access/download;Figure;Fig.2.tiff



Sample type	Measurement	Criterion
Collagen (originating from bone, tooth dentine, tendons, fish scales)	$\delta^{13}C$, $\delta^{15}N$	**%C,** calculated as amount (in µg) of C / amount (in µg) of combusted sample
	$\delta^{13}C$, $\delta^{15}N$	**%N,** calculated as amount (in µg) of N / amount (in µg) of combusted sample
	$\delta^{13}C$, $\delta^{15}N$	**C:N** atomic ratios, calculated as ((%C/12)/(%N/14))
	$\delta^{13}C$, $\delta^{15}N$	Correlation in $\boldsymbol{\delta^{13}C}$ **vs. C:N**
	$\delta^{13}C$, $\delta^{15}N$	**%collagen yield**, calculated as amount (in mg) of extracted collagen / amount of bone prior to demineralization
Charred crop remains	$\delta^{13}C$, $\delta^{15}N$	**Initial assessment of charring temperature** using both surface morphology and internal grain structure (cross-sections)
	$\delta^{13}C$, $\delta^{15}N$	**FTIR** to rule out contamination by carbonates and nitrates
	$\delta^{13}C$, $\delta^{15}N$	Graph $\boldsymbol{\delta^{13}C}$ **vs C:N**, $\boldsymbol{\delta^{13}C}$ **vs %C**, $\boldsymbol{\delta^{15}N}$ **vs C:N**, $\boldsymbol{\delta^{15}N}$ **vs %N**. Any correlations signal diagenetic alteration.
Tooth enamel bioapatite	$\delta^{13}C$, $\delta^{18}O$	**Crystallinity index**, calculated as $(A_{565} + A_{605})/A_{595}$, where A is the absorbance at the wavelength specified by the subscript
		FTIR to assess presence of calcite, francolite, and amorphous carbonate
		Correlation between $\delta^{18}O_{carbonate}$ vs. $\delta^{18}O_{phosphate.}$
	$\delta^{18}O$	$\boldsymbol{CO_3}$**:**$\boldsymbol{PO_4}$ **ratio** calculated as the percent amount of CO3/percent amount of PO4

		BPI (B-carbonate to Phosphate Index) of ~0.6 ± 0.1
Bone bioapatite	$\delta^{13}C$, $\delta^{18}O$	**Bioapatite yield**
		C/P calculated using **FTIR–ATR** or **DRIFT**
		IR-SF calculated using **FTIR–ATR** or **DRIFT**
Molluscan shell	$\delta^{13}C$, $\delta^{18}O$	Visual examination by **microscopy** to examine preservation of carbonate structure.
		FTIR-ATR to discriminate aragonitic (original) vs calcitic (recrystallized) carbonate amongst certain species.
		Analytical chemistry methods such as **raman spectroscopy** and **XRD**.

Current benchmarks
> 35 ‰ by weight
> 11–16% by weight
For modern/historical collagen : C:N ratios between 3.00–3.30 for fish and 3.00–3.28 for mammals and birds. *For archaeological collagen* : C:N ratios vary based on context but range from: 1) 2.90 to between 3.40 and 3.90 for mammals, birds, and warm water fish; or, 2) 2.90 to between 3.20 and 3.55 for cold water fish.
Correlation signals diagenetic alteration. Requires larger datasets and is more appropriate for ecologically similar or single-species data that are from the same/similar geographic
above ~0.5–2.0‰; although the yield can vary depending on taxon and extraction methods
Only seeds assessed to be within a charring window of c.215–300 °C should be analyzed isotopically.
Carbonate peaks at 720 cm^{-1} and 870 cm^{-1}; Nitrate peaks at 1085 cm^{-1}, 1450 cm^{-1}, 3300 cm^{-1}
Correlation in any of these relations signal diagenetic alteration.
Samples with CI within 2.9–3.3 are unlikely to have recrystallized
Calcite peaks at 1407cm^{-1}, 875cm^{-1} and single peak at 711cm^{-1}; francolite at 1096 cm^{-1}; amorphous carbonate content at 866 cm^{-1};
The slope of the correlation can be used to assess for inorganic oxygen exchange and bacterially-mediated diagenesis.
A ratio of ~0.2 suggests that enamel sample is unlikely to have incorporated exogenous carbonate or lost structural carbonate

A BPI of ~0.6 ± 0.1 indicates that the sample did not undergo significant chemical alteration
Yield between 29–62% suggests that the samples have not been chemically altered
C/P values between 0.08 and 0.18 suggest that the samples have not been chemically altered
IR-SF values between 3.13–3.89 suggest that the samples have not been chemically altered
E.g. chalkiness or flakiness indicates recrystallisation, inspect definition of growth
E.g. calcite peaks at c. 1407cm^{-1}, 875cm^{-1} and single peak at 711cm^{-1}. Aragonite peaks at c. 1454cm^{-1}, 856cm^{-1} and double peak at 700cm^{-1} and 711cm^{-1}
Benchmarks differ according to method. Analyses may be relatively expensive and excessively destructive so are frequently applied to only a small number of shells, and results often not reported fully.

Reference
van Klinken (1999)
van Klinken (1999)
Originally DeNiro (1985), but more recently Guiry and Szpak (2020)
Guiry and Szpak (2021)
Ambrose (1990), DeNiro and Weiner (1988), van Klinken (1999), Dobberstein et al. (2009), Guiry et al. (2016)
Charles et al. (2015), Nitsch et al. (2015) Stroud et al. (in prep)
Vaiglova et al. (2014)
e.g. Szpak and Chiou (2019)
Weiner and Bar-Yosef (1990), Shemes (1990), Reyes-Gasga et al. (2013)
Zazzo et al. (2004), France et al. (2020)
Zazzo et al. (2004)
Wright and Schwarts (1996), Smith et al. (2007), Webb et al. (2014)

Webb et al. (2014)
Chesson et al. (2021)
France et al. (2020)
France et al. (2020)
Aldeias et al. (2019)
Milano et al. (2016)
Loftus et al. (2015)
e.g. Yanes et al. (2011, 2014)

Click here to access/download;Table;Table 2.xlsx

Species	Difference in $\delta^{13}C$ values (‰)	Confidence Interval	p-value
Pea – free-threshing wheat	–2.3	(–3.2, –1.5)	0
Hulled barley – free-threshing wheat	–1.6	(–2.3, –0.8)	<0.001
Pea – einkorn	–1.4	(–2.5, –0.4)	0.0039
Pea – hulled barley	–0.7	(–1.6, 0.1)	0.107
Hulled barley – einkorn	–0.7	(–1.6, 0.2)	0.189
Free-threshing wheat – einkorn	0.9	(0.0, 1.8)	0.056

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

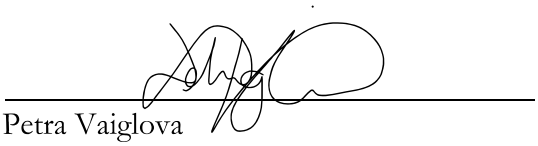
☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Quaternary International

We the authors declare that this manuscript is original, has not been published before and is not currently being considered for publication elsewhere.
We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.
We understand that the Corresponding Author is the sole contact for the Editorial process. He/She is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs.

Sincerely,

On behalf of all authors



Petra Vaiglova

Click here to access/download
E-component/Supplementary Material
Vaiglova et al. SM 1.docx

Click here to access/download
E-component/Supplementary Material
Vaiglova et al. SM 2.docx

Click here to access/download
E-component/Supplementary Material
Vaiglova et al. SM 3.docx