Article Type: Method Paper

Corresponding author mail id: gsharman@uark.edu

detritalPy: A Python-based Toolset for Visualizing and Analyzing Detrital Geo-Thermochronologic Data

Glenn R. Sharman[1], Jonathan P. Sharman[2], and Zoltan Sylvester[3]

[1] Department of Geosciences, University of Arkansas, Fayetteville, AR

[2] Department of Computer Science, Rice University, Houston, TX

[3] Bureau of Economic Geology, Jackson School of Geosciences, The University of Texas at Austin, Austin, TX

ABSTRACT

Detrital geochronology and thermochronology have emerged as primary methods of reconstructing the tectonic and surficial evolution of the Earth over geologic time. Technological improvements in the acquisition of detrital geo-thermochronologic data have resulted in a rapid increase in the quantity of published data over the past two decades, particularly for the mineral zircon. However, existing tools for visualizing and analyzing detrital geo-thermochronologic data generally lack flexibility for working with large datasets, hampering efforts to utilize the large quantity of available data.

This paper presents detritalPy, a Python-based toolset that is designed for flexibility in visualizing and analyzing large detrital geo-thermochronologic datasets. Any number of samples, or groups of samples, can be selected for plotting and/or analysis. Functionality includes: (1) plotting detrital age distributions using the most commonly employed

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/dep2.45

visualization types, (2) plotting sample locations within an interactive mapping interface, (3) calculating and plotting maximum depositional age, (4) creating multi-dimensional scaling plots, and (5) calculating inter-sample similarity and dissimilarity matrices, among other functions. detritalPy is implemented using a Jupyter Notebook, requires no significant coding expertise, and can be modified as needed to meet users' specific requirements. It is anticipated that detritalPy will provide a platform for analyzing detrital geo-thermochronologic data within a 'Big Data' framework, providing a much needed toolset for efficient utilization of ever-increasing quantities of data.

Keywords

Detrital zircon, geochronology, Python, thermochronology.

INTRODUCTION

The fields of detrital geochronology and thermochronology aim to identify the timing of crystallization and cooling of individual detrital mineral grains, respectively, with application to understanding both solid-earth (tectonic) and earth surface processes (Fedo *et al*., 2003; Reiners & Brandon, 2006; Gehrels, 2011). Technological improvements over the last several decades, particularly in laser ablation-inductively coupled plasma-mass spectrometry (LA-ICP-MS), have led to greater efficiency in collecting detrital geochronologic data (Gehrels, 2014), and to a lesser extent thermochronologic data (Horne *et al*., 2016). Although many detrital mineral types are used in geo-thermochronology, U-Pb and (U-Th)/He dating of detrital zircon (DZ) has emerged as a primary tool for constraining sediment provenance, refining palaeogeographic and tectonic interpretations, and constraining the depositional age of stratigraphic sequences (Fedo *et al*., 2003; Gehrels, 2014). Termed the "DZ Revolution" (Gehrels, 2011), the number of scientific articles, including peer-reviewed publications and conference abstracts, published per year that contain the phrase "detrital zircon" in the title has increased from ~14 per year during the early 1990's to nearly 600 in 2016 (Fig. 1A). During this time period, both the number of grain analyses per sample and the total number of samples per study has also increased (Fig. 1A), resulting in an ever-expanding number of geo-thermochronologic data points.

How many individual DZ geochronologic analyses have been collected within the past 20 years? Voice *et al*. (2011) and Puetz *et al*. (2018) compiled ~200,000 and ~260,000 individual DZ U-Pb ages, respectively, but these compilations are likely a small fraction of the total that has been collected. Back-of-the-envelope calculations suggest a conservative estimate of several million or more published U-Pb analyses of DZ alone (Fig. 1B). This is likely a conservative estimate as not all articles that publish detrital geochronologic data contain the exact phrase "detrital zircon" in their title, and this estimate does not include unpublished data. The trend towards a large-*n* sampling strategy (Pullen *et al*., 2014; Daniels *et al*., 2017) suggests that the quantity of detrital geochronologic data will continue to increase at a rapid pace. Although detrital thermochronologic datasets are typically smaller than their geochronologic counterparts, as a result of greater effort and expense of data collection, ongoing advances in He-dating using laser ablation shows promise in increasing the future quantity of detrital thermochronometric data (Horne *et al*., 2016).

The proliferation of detrital geo-thermochronologic data provides an opportunity for researchers to leverage published data in interpreting their own datasets (Gehrels, 2014). Yet efficient management and analysis of such quantities of data can make even common tasks difficult, such as (1) selecting samples for comparison, (2) combining samples into groups, and (3) assessing the similarity and/or dissimilarity between samples or groups of samples (e.g., multi-dimensional scaling; Vermeesch, 2013). A number of analytical and visualization tools have been developed for working with detrital geo-thermochronologic data, including Isoplot (Ludwig, 2008), Excel-based macros from the Arizona LaserChron Center, DZStats (Saylor & Sundell, 2016), DZMix (Sundell & Saylor, 2017), and a number of tools developed by Pieter Vermeesch and his colleagues including Density Plotter (Vermeesch, 2012), MuDiSc (Vermeesch, 2013), and the R provenance package (Vermeesch *et al*., 2016).

detritalPy, a Python 3-based toolset presented herein, supplements these existing tools by allowing efficient visualization and analysis of large detrital geochronologic and thermochronologic datasets. The following sections provide an overview of the data format required by detritalPy and an explanation of its visualization and analysis tools. Additional explanation is provided in the detritalPy manual (see supporting information) and within commented lines within the Python code itself (https://github.com/grsharman/detritalPy). Example datasets from Sharman *et al*. (2013), Sharman *et al*. (2015), and Thompson *et al*. (2017) are used to illustrate detritalPy functions.

DATA STRUCTURE

detritalPy requires input data to be structured by sample and detrital analysis (Fig. 2). Samples must contain a unique alphanumeric identifier (e.g., "11-Escanilla"; Fig. 2) and each sample must have at least one detrital analysis. The default data input format is a Microsoft Excel spreadsheet with two worksheets. (1) A "Samples" worksheet contains a row for each unique sample in the dataset. This worksheet can optionally contain other sample information (e.g., latitude and longitude coordinates). (2) An "ZrUPb" worksheet contains a row for each unique detrital analysis in the dataset, in this case zircon U-Pb and (U-Th)/He ages. Each analysis must be linked with a sample by its unique sample identifier (i.e., "11-Escanilla", Fig. 2). Each analysis must have at least one detrital age with an associated analytical uncertainty (1- or 2-sigma). If multiple detrital analyses are from the same grain (e.g., rim and core analyses), then a unique grain identifier can be used (e.g., "7_Guaso_81"; Fig. 2B). The "ZrUPb" worksheet can optionally contain other information related to the detrital analysis (e.g., U concentration or Th/U).

DATA LOADING and SAMPLE SELECTION

Data can be loaded into detritalPy by simply specifying the file pathway and file name and extension (Fig. 3A). If data is present in multiple spreadsheets, these can be imported simultaneously and will be merged together, provided that there is no duplication of data and that both spreadsheets use the same column heading names. A histogram of the number of grains per sample can be optionally plotted (Fig. 3A).

Samples can be selected for plotting and/or analysis via one of two options: (1) one or more individual samples can be specified by listing each unique sample identifier in an array, (2) one or more groups of samples can be specified by listing the sample identifiers that make up each group in an array, followed by an alphanumeric name for the group, all contained within a tuple data structure (Fig. 3B).

DETRITALPY FUNCTIONS

The visualization and analysis functions included with detritalPy are described below. All functions can be modified to fit the user's requirements by modifying the source code within the included detritalFuncs library. Additional information for each of the following functions is provided within the detritalPy manual (see supporting information) and within commented lines within the Python code itself.

Plot detrital age distributions

Detrital age distributions can be plotted for individual samples or sample groups using the most popular visualization types, including cumulative distributions, relative probability distributions, histograms, and pie diagrams (Fig. 4). There is no limit to the number of samples or sample groups that can be plotted. The plot is divided into two parts. (1) An upper subplot contains superimposed cumulative distributions for each sample or sample group. (2) One or more lower subplot(s) includes relative probability distributions, histograms, and pie diagrams, if selected to be plotted. Each sample or sample group will be plotted in the order that the unique sample identifiers are listed. If the 'separateSubplots' variable is set to True, then each sample or sample group will be plotted in a separate subplot (Fig. 5). The y-axes of the subplots will have the same scale (i.e., normalized) if the 'normPlots' variable equals True (Fig. 5). If the 'separateSubplots' variable equals False, then each normalized distribution (for a sample or sample group) will be vertically stacked within a single subplot (Fig. 5). However, histograms and pie diagrams cannot be plotted if 'separateSubplots' equals False.

Two types of relative probability distributions can be plotted. Probability density plots (PDPs) are probability distributions constructed from summing a Gaussian distribution for each analysis and normalizing the distribution such that its integral equals 1, where the mean and standard deviation of each summed Gaussian distribution is equal to the age and 1σ analytical uncertainty of the analysis (Vermeesch, 2012). Kernal density estimations (KDEs) are constructed in a similar manner to PDPs, except that a bandwidth is used for the standard deviation of each Gaussian, rather than the 1σ analytical uncertainty (Vermeesch, 2012). Although the PDP lacks a solid theoretical foundation and has been criticized as a visualization approach for detrital age distributions (Vermeesch, 2012, 2018), this option is

included as the PDP continues to be widely used by the detrital geochronology community. Cumulative PDPs and/or KDEs can be also be plotted (e.g., Fig. 4B).

Three options for colouring relative probability distributions have been included: (1) solid colouration that fills the area under the age distribution and that matches the line colour of the cumulative distribution, if selected for plotting (e.g., Fig. 6), (2) age-dependent colouration that fills the area under the relative age distribution and that correspond to user-defined detrital age categories (e.g., Fig. 4F), and (3) vertical, coloured bars that correspond to user-defined age categories. The user-specified age categories and colours can also be used to plot pie diagrams (Fig. 4F).

Plot rim age versus core age

Some detrital grains contain younger mineral growth (i.e., a rim) over an older core. Rim versus core age relationships can be plotted provided that (1) grains are identified with a unique alphanumeric label (e.g., "7_Guaso_81"), and (2) "Rim" and "Core" designations are included in a "RimCore" column (Fig. 2B). Figure 7 presents an example plot of rim versus core age from the Ainsa Basin of the Spanish Pyrenees (data from Thompson *et al*., 2017). Data points are automatically coloured according to either the sample or sample group, and error bars can be plotted, optionally.

Plot detrital age distributions in comparison to another variable (e.g., Th/U)

Geochemical attributes of detrital minerals can provide additional insight into sedimentary provenance and/or the petrogenesis of source rocks (Barth *et al*., 2013; Malkowski & Hampton, 2014; Colgan & Stanley, 2015). For detrital zircon, concentrations of U and Th are routinely measured in LA-ICP-MS. Increasingly, the abundances of other trace and rare-earth elements and Hf isotopes are also analyzed (Gehrels, 2014). Detrital U-Pb ages can be plotted in comparison to any other numeric variable in the "ZrUPb" worksheet (e.g., the concentration of U or ratio of Th to U; Fig. 2B) to assess changes in this variable with respect to the age of the detrital mineral (Fig. 8). Error bars can be plotted optionally; error can either be specified as a percentage or as the column heading that contains the variable error data. The option is also provided to plot a moving average of a specified window size (Fig. 8).

Plot detrital age populations as a bar graph

Bar graphs can provide a useful visualization of how detrital ages vary between samples or sample groups (Sharman *et al*., 2015; their Fig. 7), with the caveat that such plots involve subjective selection of bin boundaries that may over-simplify complex age distributions. detritalPy allows age proportions to be plotted as one or more bar graphs, using user-specified age categories and colours (Fig. 9). If plotting sample groups, setting the variable 'separateGroups' to True results in plot(s) with the age proportions for individual samples in each group (Fig. 9B). Otherwise, the proportions will reflect all analyses within each group (Fig. 9C).

Plot sample locations on an interactive map

Samples with latitude and longitude coordinates can be plotted on an interactive map, provided that the folium library has been installed (https://github.com/python-visualization/folium; Fig. 10). A number of basemap options are available and can be specified through the variable 'mapType'. Age distributions pop-ups (e.g., PDP or KDE) can be enabled and viewed interactively by clicking on each sample. Sample locations, including the unique sample identifier and an optional descriptor (e.g., the 'Unit' category; Fig. 2A), can also be exported as a Google Earth kml file.

Plot and export maximum depositional age (MDA) calculations

The youngest DZ U-Pb ages provide an estimate of the maximum depositional age (MDA) of a detrital sample (Fedo *et al*., 2003; Dickinson & Geherls., 2009). An automated approach is provided to calculating the MDA of one or more sample(s) or sample group(s) using three *ad hoc* metrics used by Dickinson & Gehrels (2009). (1) The youngest single grain, YSG, assigns the MDA as the youngest detrital analysis (Fig. 11). The YSG is defined by sorting all analyses by their U-Pb age plus 1σ uncertainty, and selecting the first analysis. Thus it is possible to have a younger, but less precise, age than the YSG as defined herein. (2) The youngest cluster of 2 or more ages with overlapping 1σ uncertainties, YC1σ(2+), has the advantage of not relying on a single analysis that could be affected by lead loss or other analytical problems (Dickinson & Gehrels, 2009). The YC1σ(2+) is defined by sorting all

analyses by their U-Pb age plus 1σ uncertainty, and identifying the youngest cluster of analyses with overlapping 1σ error (Fig. 11). (3) The youngest cluster of 3 or more ages with overlapping 2σ uncertainties, YC2σ(3+), provides a more conservative, but typically older, estimate of the MDA than the other two metrics (Dickinson & Gehrels, 2009). The YC2σ(3+) is defined by sorting all analyses by their U-Pb age plus 2σ uncertainty, and identifying the youngest cluster of 3 or more analyses with overlapping 2σ error (Fig. 11). A spreadsheet with all MDA calculation results is exported automatically.

The MDA calculations can be optionally plotted for each sample or sample group selected (Fig. 11). Only the grain ages that are used in at least one of the three MDA calculations will be included. Analyses can be arranged by their mean age, mean age plus 1-sigma analytical uncertainty, or mean age plus 2-sigma analytical uncertainty. Additional plot parameters can be used to change the appearance of the plot, including its dimensions, the width of the bars, and colours to use for the different MDA calculations.

Multi-dimensional scaling

Multi-dimensional scaling (MDS) has become a popular approach for visual assessment of the degree of similarity and dissimilarity between detrital geochronologic samples (Vermeesch, 2013; Saylor *et al*., 2017). MDS plots for samples or sample groups can be created in detritalPy using the sklearn library with the option for either metric or non-metric MDS (Fig. 12). Following Vermeesch (2013) and Saylor *et al*. (2017), MDS calculations can be based on the maximum separation between cumulative distribution functions (CDF) (Kolgomorov-Smirnov D_{max}) or the sum of the maximum differences between two CDFs (i.e., CDF_1-CDF_2 and CDF_2-CDF_1; Kupier V_{max}). The option is provided to plot data points as pie diagrams, using user-specified age categories and colours. Visualizing MDS plots using pie diagrams can help with interpreting the spatial distribution of samples on an MDS plot (Fig. 12).

(U-Th)/He vs U-Pb age "double dating" plot

Grains with both U-Pb crystallization and (U-Th)/He cooling ages can be plotted against each other in a 'double dating' plot (Thompson *et al*., 2017). The main portion of the plot contains a scatter plot, with gray shading within the region where the cooling age is older than the crystallization age (Fig. 13). Separate subplots on the x- and y-axis show the relative probability distribution (PDP and/or KDE) of the U-Pb and (U-Th)/He age distributions, respectively (Fig. 13). Histograms can be plotted, optionally.

Export sample comparison matrices as a CSV file

A number of metrics have been proposed to evaluate the similarity or dissimilarity between detrital age distributions (Saylor & Sundell, 2016; and references within), although the use of some metrics (e.g., likeness, cross-correlation) has recently been discouraged (Vermeesch, 2018). Sample comparison matrices for samples or sample groups can be exported to a CSV file, including similarity, likeness, the Kolgomorov-Smirnov statistic D_{max} and p-value, the Kuiper statistic V_{max}, and the cross-correlation (r^2) coefficient of either the PDP or KDE. Similarity, likeness, and r^2 coefficient values are based on selection of either the PDP or KDE distribution, and when based upon the KDE, will depend on choice of a bandwidth.

Export detrital age distributions as a CSV file

Raw detrital age distributions (either cumulative or relative) can be exported as a CSV file. The age range and discretization interval can be specified. If the variable 'normalize' equals True, the distribution(s) will be forced to sum to 1. Exported age distributions can be used as inputs for more advanced analytical procedures (e.g., sediment unmixing; Sharman & Johnstone, 2017).

Export ages and errors in tabular format as a CSV file

To promote compatibility with other software, detritalPy includes the option to export a CSV file containing U-Pb ages and 1σ uncertainties for any number of samples or sample groups. U-Pb ages are automatically sorted from youngest to oldest, and listed in adjacent columns, the format required by many existing tools (e.g., Arizona LaserChron Center Excel worksheets).

DISCUSSION

KDE Bandwidth Selection

Selection of an appropriate KDE bandwidth is important for avoiding over-smoothing or under-smoothing detrital age distributions (Vermeesch, 2012; Saylor & Sundell, 2016). detritalPy offers three options for KDE bandwidth selection. The bandwidth can be specified by the user in units of Myr. Alternatively, the bandwidth can be automatically selected by an algorithm that attempts to select an optimized value (Shimazaki & Shinomoto, 2010) using either a fixed or variable bandwidth (Fig. 14), as implemented through the adaptivekde library (https://pypi.python.org/pypi/adaptivekde).

To consider the influence of KDE bandwidth selection on the appearance of detrital age distributions, two end-member cases are considered: a single sample from the Morrison Formation of central Colorado (149 analyses; Sharman *et al*., *in press*) and a large compilation of 1,308 samples (120,968 analyses) that are mostly from North American continent (Fig. 14). Application of the Shimazaki and Shinomoto (2010) algorithm to the single sample yields an optimized, fixed bandwidth of 17.8 Myr. This bandwidth selection appears to yield an acceptable match with the histogram over much of the age range of the detrital analyses, but over-smooths the youngest, Jurassic age peak (Fig. 14A). Use of a narrower bandwidth (e.g., 5 Myr) results in better reproduction of the Jurassic age peak, and greater similarity to the PDP, but appears to under-smooth the older Palaeozoic and Proterozoic age peaks (Fig. 14A). Although the optimized, variable bandwidth (Shimazaki & Shinomoto, 2010) is able to better reproduce the precision of the young, Jurassic peak, it appears to over-smooth the older age components. For the large data compilation, the differences in appearance between the PDP and different KDE bandwidths are less than with the single sample (Fig. 14B). The optimized, fixed bandwidth (2.1 Myr) yields a result that

closely resembles the PDP, 5 Myr bandwidth, and optimized, variable bandwidth plots (Fig. 14B). However, the 10 Myr and 20 Myr bandwidths appear to over-smooth the Mesozoic-Cenozoic age populations (Fig. 14B).

It is suggested here that the selection of a KDE bandwidth, whether user-defined or based on an optimized routine (Botev *et al*., 2010; Shimazaki & Shinomoto, 2010), will ultimately depend upon the nature of the age distributions themselves and the intended purpose of the display, with the decision having the greatest impact on plotted distributions with low numbers of analyses. Plots comprised of a relatively small number of detrital analyses may typically require larger bandwidths to avoid under-smoothing regions of sparse data while also tending to over-smooth young, precise age peaks (Fig. 14A). This issue may be partially alleviated by plotting young and old detrital analyses separately, using different bandwidths for each (Sharman *et al*., *in press*). Plots comprised of large numbers of analyses, however, may benefit from selection of a smaller bandwidth (Fig. 14B). Note that the choice of a KDE bandwidth has limited influence on the appearance of the cumulative (summed) KDE relative to the significant impact on the appearance of the KDE (Fig. 14).

Strengths of detritalPy

Although intended as a complement to rather than replacement of existing tools, detritalPy has a number of strengths that allow for efficient visualization and analysis of large detrital geo-thermochronologic datasets:

(1) The code is executed in the open-source (Python Software Foundation License) Python 3 language and is implemented using a user-friendly Jupyter Notebook interface (Perez and Granger, 2007; Kluyver *et al*., 2016). Thus, detritalPy does not require the use of proprietary software (e.g., MathWorks Matlab). Although no significant coding knowledge is required, users have the option of modifying the code to create user-customized functions and plots, which can be difficult with some existing tools that utilize a graphical user interface.

(2) All detritalPy functions are compatible with an unlimited number of samples, or sample groups, and samples can be selected without manipulation of data in spreadsheets. Samples can be selected via simple reference to the unique sample identifier (e.g., 'Sample B'), and changes can be made on-the-fly. Thus detritalPy eliminates the need to manually

combine or organize data prior to plotting or analysis, beyond initial data organization (Fig. 2), helping to eliminate duplication of data in separate spreadsheets.

(3) Plotting and analysis functions are designed to allow maximum user flexibility in controlling the appearance and types of plots, following the most commonly used visualization and analytical approaches (Vermeech, 2012; 2013; Saylor & Sundell, 2016). Jupyter Notebooks are ideal for exploratory data analysis. Once the desired graphs have been created, plots can be can be exported in a vector-friendly format, requiring little modification for publication-quality figures.

(4) Data can be exported in the common format used by the majority of existing analytical and visualization tools, for use in other published software.

'Big Data' in Detrital Geochronology

The proliferation of detrital geo-thermochronologic data within the last 20 years (Fig. 1) provides an opportunity for analysis within a 'Big Data' framework (Vermeesch & Garzanti, 2015). Yet development of an efficient means of visualizing and analyzing large datasets remains a critical need in the detrital geo-thermochronologic community (Geherls, 2014). For example, the ability to easily create sample groups within detritalPy will facilitate construction of reference curves (Gehrels *et al*., 1995; Kimbrough *et al*., 2015) that can be compared with the magmatic and/or metamorphic history of known basement terranes (Dickinson & Gehrels, 2009) or with other detrital samples (Sharman *et al*., 2017). The ability to quickly visualize and analyze related detrital geochronological and geochemical data also has relevance for characterization of source terranes, such as the magmatic flux of volcanic arc terranes (Ducea, 2001; Sharman *et al*., 2015; Malkowski *et al*., 2017).

detritalPy also provides improved functionality for querying and exploring geo-thermochronologic data, particularly when combined with existing data repositories (e.g., the Geochron database; www.geochron.org). The ability to plot sample locations on an interactive, zoomable map and export them to Google Earth allows rapid identification of geographically related samples (Fig. 10). Relationships between samples or sample groups can be quickly assessed both visually and quantitatively (e.g., Figs 6 and 12). For instance, plotting samples or sample groups in stratigraphic succession (Gehrels *et al*., 2011) has potential for assessing the degree of multi-generational sediment recycling over time

(Thomas, 2011). Because detritalPy is open-source, its functions can be modified or added to, as needed. It is anticipated that future development of new visualization and analytical tools within the detritalPy framework will address the evolving needs of the detrital geo-thermochronologic community.

CONCLUSIONS

detritalPy, a Python-based approach to visualizing and analyzing large detrital geo-thermochron datasets, addresses a critical need for an efficient means of processing the rapidly expanding quantity of detrital mineral isotopic and geochemical data. detritalPy is implemented through a user-friendly Jupyter Notebook interface and requires no significant coding expertise. However, the existing code can be modified to allow for user-customized plots and analysis.

An unlimited number of samples can be either plotted individually or within groups. Functionality includes (1) plotting detrital U-Pb age distributions using the most commonly employed visualization types, (2) plotting rim age versus core age, (3) comparing detrital age distributions to another variable (e.g., Th/U), (4) plotting age group proportions as a bar graph, (5) plotting sample locations on an interactive, zoomable map and exporting a Google Earth kml file, (6) calculating and visualizing maximum depositional ages, (7) multi-dimensional scaling, (8) calculation of similarity and dissimilarity metrics (e.g., similarity, likeness, Kolgomorov-Smirnov statistic), and (9) exporting U-Pb age and error data and age distributions as CSV files.

detritalPy has a number of advantages over existing tools, including not requiring proprietary software, offering flexibility in how data are plotted and analyzed, and eliminating the need to manipulate data within spreadsheets to select which samples or groups of samples to plot. Furthermore, data can be easily exported from detritalPy in the common format required by most existing data visualization and analytical tools. It is expected that detritalPy will provide an important toolset for analyzing detrital geochronologic and thermochronologic data within a 'Big Data' framework.

ACKNOWLEDGMENTS

We would like to acknowledge coding assistance from Kevin Befus and Samuel Johnstone. Support for this research was provided by the industrial affiliate members of the Quantitative Clastics Laboratory at the Bureau of Economic Geology. We thank Joel Saylor and Pieter Vermeesch for constructive feedback in peer-review.

REFERENCES

Barth, A.P., Wooden, J.L., Jacobson, C.E. and **Economos, R.C.** (2013) Detrital zircon as a proxy for tracking the magmatic arc system: The California arc example: *Geology*, **41**, 223-226.

Botev, Z.I., Grotowski, J.F. and **Kroese, D.P.** (2010) Kernal density estimation via diffusion: *Annals of Statistics*, **38**, 2916-2957.

Colgan, J.P. and **Stanley, R.G.** (2015) The Point Sal-Point Pedras Blancas correlation and the problem of slip on the San Gregorio-Hosgri fault, central California Coast Ranges: *Geosphere*, **12**, 971-984.

Daniels, B.G., Auchter, N.C., Hubbard, S.M., Romans, B.W., Matthews, W.A. and **Stright, L.** (2017) Timing of deep-water slope evolution constrained by large-*n* detrital and volcanic ash zircon geochronology, Cretaceous Magallanes Basin, Chile: *Geological Society of America Bulletin*, **130**, 438-454.

Dickinson, W.R. and **Gehrels, G.E.** (2009) Use of U-Pb ages of detrital zircons to infer maximum depositional ages of strata: A test against a Colorado Plateau Mesozoic database: *Earth and Planetary Science Letters*, **288**, 115-125.

Ducea, M. (2001) The California arc: Thick granitic batholiths, eclogitic residues, lithospheric-scale thrusting, and magmatic flare-ups: *GSA Today*, **11**, 4-10.

Fedo, C.M., Sircombe, K.N. and **Rainbird, R.H.** (2003) Detrital zircon analysis of the sedimentary record: *Reviews in Mineralogy and Geochemistry*, **53**, 277-303.

Gehrels, G.E. (2014) Detrital zircon U-Pb geochronology applied to tectonics: *Annual Review of Earth and Planetary Sciences*, 42, 127-149.

Gehrels, G.E., Blakey, R., Karlstrom, K.E., Timmons, J.M., Dickinson, W.R. and **Pecha, M.** (2011) Detrital zircon U-Pb geochronology of Paleozoic strata in the Grand Canyon, Arizona: *Lithosphere*, 3, 183-200.

Gehrels, G.E., Dickinson, W.R., Ross, G.M., Stewart, J.H. and **Howell, D.G.** (1995) Detrital zircon reference for Cambrian to Triassic miogeoclinal strata of western North America: *Geology*, **23**, 831-834.

Horne, A.M., van Soest, M.C., Hodges, K.V., Tripathy-Lang, A. and **Hourigan, J.K.** (2016) Integrated single crystal laser ablation U/Pb and (U-Th)/He dating of detrital

accessory minerals – Proof-of-concept studies of titanites and zircons from the Fish Canyon tuff: *Geochimica et Cosmochimica Acta*, **178**, 106-123.

Kimbrough, D.L., Grove, M., Gehrels, G.E., Dorsey, R.J., Howard, K.A., Lovera, O., Aslan, A., House, K. and **Pearthree, P.A.** (2015) Detrital zircon U-Pb provenance of the Colorado River: A 5 m.y. record of incision into cover strata overlying the Colorado Plateau and adjacent regions: *Geosphere*, **11**, 1719-1748.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C. and **Jupyter Development Team** (2016) Jupyter Notebooks – a publishing format for reproducible computational workflows. In: *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (Eds. Loizides, F. and Schmidt, B.), pp. 87-90. IOS Press, Amsterdam, Netherlands, doi:10.3233/978-1-61499-649-1-87.

Ludwig, K.R. (2008) Isoplot/Ex version 3.7: A geochronologic toolkit for Microsoft Excel: *Berkeley Geochronologic Center Special Publication*, **4**, 1-77.

Malkowski, M.A. and **Hampton, B.A.** (2014) Sedimentology, U-Pb detrital geochronology, and Hf isotopic analyses from Mississippian-Permian stratigraphy of the Mystic subterrane, Farewell terrane, Alaska: *Lithosphere*, **6**, 383-398.

Malkowski, M.A., Schwartz, T.M., Sharman, G.R., Sickmann, Z.T. and **Graham, S.A.** (2017) Stratigraphic and provenance variations in the early evolution of the Magallanes-Austral foreland basin: Implications for the role of longitudinal versus transverse sediment dispersal during arc-continent collision: *Geological Society of America Bulletin*, **129**, 349-371.

Perez, F. and **Granger, B.E.** (2007) IPython: A System for Interactive Scientific Computing: *Computing in Science & Engineering*, **9**, 21-29.

Puetz, S.J., Ganade, C.E., Zimmermann, U. and **Borchardt, G.** (2018) Statistical analysis of Global U-Pb Database 2017: *Geoscience Frontiers*, **9**, 121-145.

Pullen, A., Ibáñez-Mejía, M., Gehrels, G., Ibáñez-Mejía, J.C. and **Pecha, M.** (2014) What happens when $n = 1000$? Creating large-n geochronological datasets with LA-ICP-MS for geologic investigations: *Journal of Analytical Atomic Spectrometry*, **29**, 971-980.

Reiners, P.W. and **Brandon, M.T.** (2006) Using thermochronology to understand orogenic erosion: *Annual Reviews in Earth and Planetary Science*, **34**, 419-466.

Saylor, J.E. and **Sundell, K.E.** (2016) Quantifying comparison of large detrital geochronology data sets: *Geosphere*, **12**, 203-220.

Saylor, J.E., Jordon, J.C., Sundell, K.E., Wang, X., Wang, S. and **Deng, T.** (2017) Topographic growth of the Jishi Shan and its impact on basin and hydrology evolution, NE Tibetan Plateau: *Basin Research*, doi:10.1111/bre.12264.

Sharman, G.R., Graham, S.A., Grove, M. and **Hourigan, J.K.** (2013) A reappraisal of the early slip history of the San Andreas fault, central California, USA: *Geology*, **41**, 727-730.

Sharman, G.R., Graham, S.A., Grove, M., Kimbrough, D.L. and **Wright, J.E.** (2015) Detrital Zircon Provenance of the Late Cretaceous-Eocene California Forearc: Influence of Laramide Low-Angle Subduction on Sediment Dispersal and Paleogeography: *Geological Society of America Bulletin*, **127**, 38-60.

Sharman, G.R., Covault, J.A., Stockli, D.F., Wroblewski, A., F.-J. and **Bush, M.A.** (2017) Early Cenozoic drainage reorganization of the U.S. Western Interior-Gulf of Mexico Sediment Routing System: *Geology*, **45**, 187-190.

Sharman, G.R. and **Johnstone, S.A.** (2017) Sediment unmixing using detrital geochronology: *Earth and Planetary Science Letters*, **477**, 183-194.

Sharman, G.R., Stockli, D.F., Flaig, P., Raynolds, R.G. and **Covault, J.A.** (*in press*) Local-to-distant provenance cyclicity of the southern Front Range, central Colorado: Insights from detrital zircon geochronology. In: *Tectonics, Sedimentary Basins and Provenance: A Celebration of the Career of William R. Dickinson* (Eds. Ingersoll, R.V., Lawton, T.F. and Graham, S.A.), Geological Society of America Special Paper 368, Boulder, CO, USA.

Shimazaki, H. and Shinomoto, S. (2010) Kernal bandwidth optimization in spike rate estimation: *Journal of Computational Neurosciences*, **29**, 171-182.

Sundell, K.E. and **Saylor, J.E.** (2017) Unmixing detrital geochronology age distributions: *Geochemistry, Geophysics, Geosystems*: doi:10.1002/2016GC006774.

Thomas, W.A. (2011) Detrital-zircon geochronology and sedimentary provenance: *Lithosphere*, **3**, 304-308.

Thomson, K.D., Stockli, D.F., Clark, J.D., Puidgefàbregas, C. and **Fildani, A.** (2017) Detrital zircon (U-Th)/(He-Pb) double-dating constraints on provenance and foreland basin evolution of the Ainsa Basin, south-central Pyrenees, Spain: *Tectonics*, **36**, doi:10.1002/2017TC004504.

Vermeesch, P. (2012) On the visualisation of detrital age distributions: *Chemical Geology*, **312-313**, 190-194.

Vermeesch, P. (2013) Multi-sample comparison of detrital age distributions: *Chemical Geology*, **341**, 140-146.

Vermeesch, P. and **Garzanti, E.** (2015) Making geologic sense of 'Big Data' in sedimentary provenance analysis: *Chemical Geology*, **409**, 20-27.

Vermeesch, P., Resentini, A. and **Garzanti, E.** (2016) An R package for statistical provenance analysis: *Sedimentary Geology*, **336**, 14-26.

Vermeesch, P. (2018) Dissimilarity measures in detrital geochronology: *Earth-Science Reviews*, **178**, 310-321.

Voice, P.J., Kowalewski, M. and **Eriksson, K.A.** (2011) Quantifying the timing and rate of crustal evolution: global compilation of radiometrically dated detrital zircons: *The Journal of Geology*, **119**, 109-126.

FIGURE CAPTIONS

Figure 1. (A) The number of published studies per year, including peer-reviewed articles and conference abstracts, that contain the phrase "detrital zircon" in the title (based on a GeoRef database search, August 2017; see also Gehrels, 2014, fig. 1). Blue and red lines demarcate the average number of samples and analyses per sample, respectively, for published studies from each calendar year (2009 data and earlier from Voice *et al*., 2011; 2010 data and later years from an unpublished compilation of ~120,000 DZ U-Pb ages, mostly from North America). (B) An approximate estimate of the total (cumulative) number of published DZ U-Pb analyses from 1990-2016. This estimate was derived from multiplying the number of published studies by the number of samples and average number of analyses per sample for each calendar year (part a).

Figure 2. Example of the default data structure used by detritalPy. (A) A "Samples" worksheet contains a required "Sample_ID" column and additional optional columns. (B) An "ZrUPb" worksheet contains required "Sample_ID", "BestAge", and "BestAge_err" columns. Additional columns are required for some plotting and analysis functions (e.g., "U_ppm"). Example data is included as supporting information.

Figure 3. (A) Example code that illustrates how to import required libraries, load data in two separate Excel files, and plot a histogram of the distribution of analyses per sample. (B) Example code that illustrates how to select one or more individual sample(s) (above) or sample group(s) (below) for plotting and analysis.

Figure 4. Examples of plotted detrital age distributions. (A) Cumulative distribution function binned at 1 Myr increments. The notation N=(X, Y/Z) indicates the number of samples (X), the number of analyses visible in the plotted age range (Y), and the total number of analyses (Z). (B) Cumulative probability density plot (CPDP). (C) Histogram with 5 Myr bins. (D) Probability density plot (PDP). (E) Kernal density estimate (KDE) using a 3 Myr bandwidth (b.w.). (F) A combination plot with a superimposed histogram, PDP, and KDE. The PDP and

pie diagram are coloured according to user-defined age categories, from Sharman *et al.* (2015).

Figure 5. Three options for plotting relative probability distributions (PDP or KDE). (A) Equally sized subplots that are allowed to have different y-axis scales. (B) Equally sized subplots that all have the same y-axis scale (i.e., normalized). (C) Distributions are stacked on top of each other and all have the same y-axis scale (i.e., normalized).

Figure 6. Detrital age distributions shown as a CKDEs (top) and KDEs (bottom) for four sample groups (using a bandwidth of 1.5 Myr). CKDE and KDE are coloured by sample group, allowing easy visual comparison of cumulative and relative age distributions.

Figure 7. Illustration of rim age versus core age relationships (data from Thompson *et al*., 2017). Data symbols are coloured by sample.

Figure 8. Illustration of plotting detrital U-Pb age distributions versus another analysis variable (Th to U ratio) for the Point of Rocks Sandstone and Butano Sandstone sample groups (see Fig. 7). The red line depicts a 15 point moving average across Th/U data points. Note elevated Th/U ratios in Jurassic zircon ages from the Butano Sandstone.

Figure 9. Bar graphs showing the relative proportions of user specified age categories (see Fig. 5). (A) Individual samples. (B) Individual samples plotted by sample group. (C) Combined data for sample groups. See Figure 7 for the Point of Rocks (POR) Sandstone and Butano Sandstone sample groups.

Figure 10. Examples of the Butano Sandstone (red) and Point of Rocks Sandstone (green) sample locations plotted on an interactive map (ESRI 'World_Topo_Map'). (A) Wide view showing all 8 samples coloured according to group. Clicking on a sample results in a pop-up window with the sample name. (B) Zoomed in view showing detail around sample BUT-5.

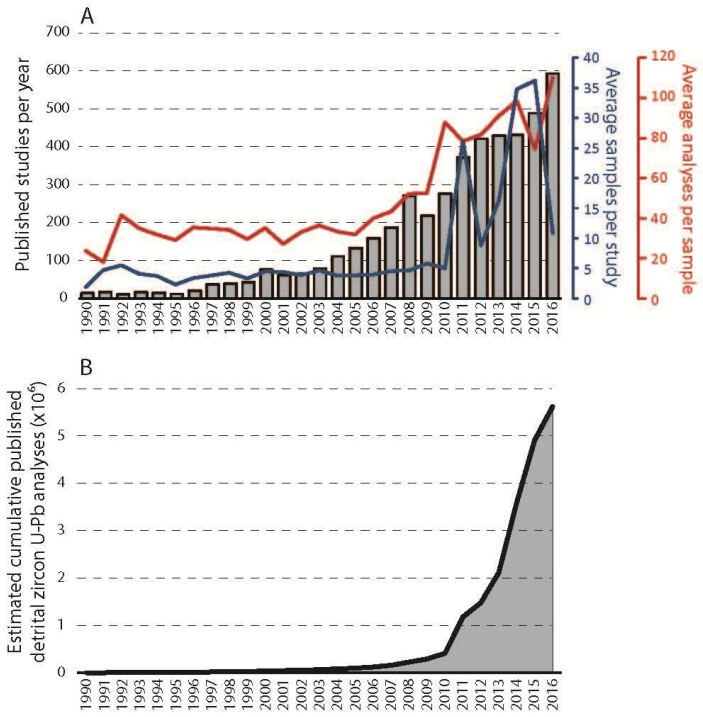
Figure 11. Example of maximum depositional age (MDA) calculations for sample POR-2. See text for details.

Figure 12. Multi-dimensional scaling (MDS) plots. Axes are dimensionless D_{max} distances (see Vermeesch, 2013 for additional explanation). (A) Individual samples. (B) Sample groups.

Figure 13. (U-Th)/He versus U-Pb 'double dating' plot. PDPs are shown in subplots to the bottom and right of the scatterplot. Data from Thompson *et al*., 2017.

Figure 14. Illustration of the influence of KDE bandwidth selection on plot appearance. (A) Single sample from the Morrison Formation of central Colorado (Sharman et al., *in press*). (B) Large data compilation of samples mostly from North America.

Sharman et al. - Figure 1



Sharman et al. - Figure 2

A

Required | Optional (not used) | Optional (used in some functions) | Optional (not used)

	A	B	C	D	E	F	G
1	Sample_ID	Unit	Basin	Age	Latitude	Longitude	Source
2	11-Escanilla	Escanilla	Ainsa Basin	Eocene (Bartonian)	42.278474	-0.122617	Thompson et al. (2017)
3	12-Escanilla	Escanilla	Ainsa Basin	Eocene (Bartonian)	42.267407	-0.116455	Thompson et al. (2017)
4	10-Sobrarbe	Sobrarbe	Ainsa Basin	Eocene (Bartonian)	42.29224	-0.101188	Thompson et al. (2017)
5	7-Guaso	Guaso	Ainsa Basin	Eocene (Lutetian)	42.409038	-0.106831	Thompson et al. (2017)
6	13-Guaso	Guaso	Ainsa Basin	Eocene (Lutetian)	42.358007	-0.156971	Thompson et al. (2017)
7	5-Morillo	Morillo	Ainsa Basin	Eocene (Lutetian)	42.379942	-0.151209	Thompson et al. (2017)
8	6-Morillo	Morillo	Ainsa Basin	Eocene (Lutetian)	42.414713	-0.11229	Thompson et al. (2017)
9	14AB-M02	Morillo	Ainsa Basin	Eocene (Lutetian)	42.43641	-0.07068	Thompson et al. (2017)
10	14AB-A04	Ainsa II	Ainsa Basin	Eocene (Lutetian)	42.433589	-0.12764	Thompson et al. (2017)
11	14AB-A05	Ainsa II	Ainsa Basin	Eocene (Lutetian)	42.43343	-0.12742	Thompson et al. (2017)
12	4-Ainsa	Ainsa I	Ainsa Basin	Eocene (Lutetian)	42.404218	-0.14801	Thompson et al. (2017)
13	14AB-A06	Ainsa I	Ainsa Basin	Eocene (Lutetian)	42.43364	-0.1314	Thompson et al. (2017)
14	15AB-352	Banaston	Ainsa Basin	Eocene (Lutetian)	42.404645	-0.190405	Thompson et al. (2017)
15	15AB-118	Banaston	Ainsa Basin	Eocene (Lutetian)	42.45504	-0.05471	Thompson et al. (2017)
16	15AB-150	Gerbe	Ainsa Basin	Eocene (Lutetian)	42.38277	-0.18547	Thompson et al. (2017)
17	3-Gerbe	Gerbe	Ainsa Basin	Eocene (Lutetian)	42.39448	-0.197896	Thompson et al. (2017)
18	14AB-G07	Gerbe	Ainsa Basin	Eocene (Lutetian)	42.39455	-0.197719	Thompson et al. (2017)
19	2-Arro	Arro	Ainsa Basin	Eocene (Ypresian)	42.406398	-0.238684	Thompson et al. (2017)
20	1-Fosado	Fosado	Ainsa Basin	Eocene (Ypresian)	42.428614	-0.256078	Thompson et al. (2017)
21	14AB-F01	Fosado	Ainsa Basin	Eocene (Ypresian)	42.434566	-0.248433	Thompson et al. (2017)

Samples | ZrUPb

B

Required | Optional (used in some functions) | Required | Optional (used in some functions)

	A	B	E	G	U	V	W	X	Y	Z
1	Sample_ID	Grain_ID	U_ppm	Th_U	BestAge	BestAge_err	Disc	ZHe_Age	ZHe_Age_err	RimCore
441	7-Guaso	7_Guaso_65	128	0.52	572	5	1.04			
442	7-Guaso	7_Guaso_60	267	0.57	575	7	1.2			
443	7-Guaso	7_Guaso_70	506	0.17	579	4.15	7.4			
444	7-Guaso	7_Guaso_81	980	0.30	590	10	1.01			Rim
445	7-Guaso	7_Guaso_86	80.5	2.63	591.4	4.9	0.94			
446	7-Guaso	7_Guaso_28	85.7	0.87	605	7	2.37			Core
447	7-Guaso	7_Guaso_92	31.28	0.64	613	6.5	1.76			
448	7-Guaso	7_Guaso_72	98.2	2.38	617.1	2.65	0.33	49.8	4.0	
449	7-Guaso	7_Guaso_49	878	0.04	624.2	4.2	1.37	202.1	16.2	
450	7-Guaso	7_Guaso_25	157.8	1.05	631.7	4.8	0.19			
451	7-Guaso	7_Guaso_53	58.3	0.88	632	6.5	1.71			
452	7-Guaso	7_Guaso_17	180.2	0.96	634.2	3.8	1.77			
453	7-Guaso	7_Guaso_46	37.3	1.25	639	5	1.39			
454	7-Guaso	7_Guaso_61	267	0.49	644.6	3.4	0.05	50.2	4.0	
455	7-Guaso	7_Guaso_7	431	0.51	645	5.5	1.23	61.5	4.9	
456	7-Guaso	7_Guaso_8	45.8	0.86	658.5	4.6	1.72			
457	7-Guaso	7_Guaso_99	105.5	0.68	684	9.5	1.01			Core
458	7-Guaso	7_Guaso_48	86.2	0.48	738	8.5	1.47			
459	7-Guaso	7_Guaso_73	75.9	0.85	742.6	4	0.79			
460	7-Guaso	7_Guaso_93	82.7	1.67	790	6	0.01			

Samples | ZrUPb

Sharman et al. - Figure 3

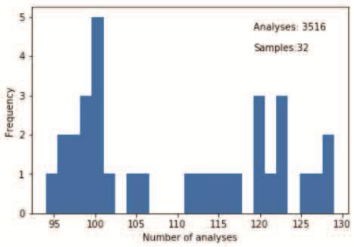
A

```
In [1]: import detritalFuncs as dFunc
        from importlib import reload
        %matplotlib inline
```

```
In [2]: # Specify file paths to data input file(s)
        dataToLoad = ['../example-data/ExampleDataset_1.xlsx',
                      '../example-data/ExampleDataset_2.xlsx']

        main_df, main_byid_df, samples_df, analyses_df = dFunc.loadData_2(dataToLoad)
```

```
In [3]: dFunc.plotSampleDist(main_byid_df, numBins=25)
```



B

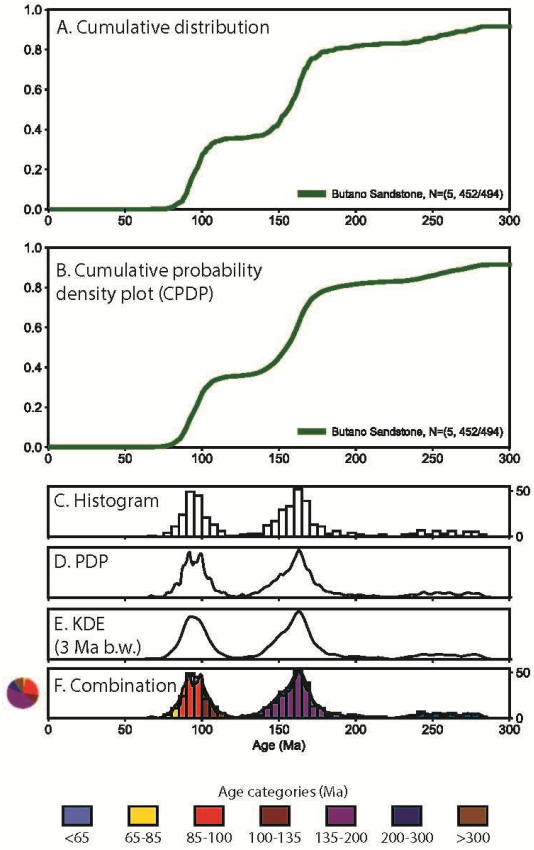
```
In [4]: sampleList = ['POR-1','POR-2','POR-3','BUT-5','BUT-4','BUT-3','BUT-2','BUT-1']

        ages, errors, numGrains, labels = dFunc.sampleToData(sampleList, main_byid_df, sigma = '1sigma');
```

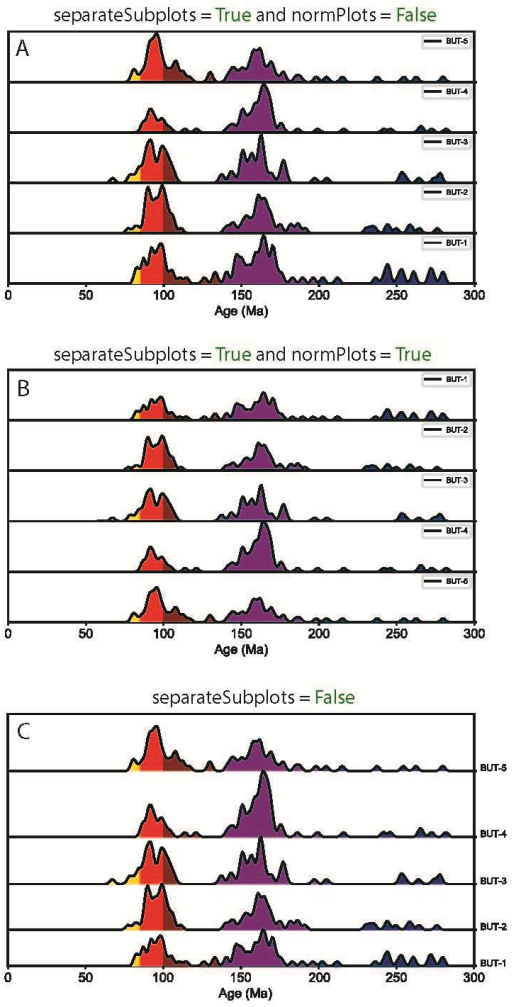
```
In [5]: sampleList = [(['POR-1','POR-2','POR-3'],'Point of Rocks Sandstone'),
                     (['BUT-5','BUT-4','BUT-3','BUT-2','BUT-1'],'Butano Sandstone')]

        ages, errors, numGrains, labels = dFunc.sampleToData(sampleList, main_byid_df, sigma = '1sigma');
```

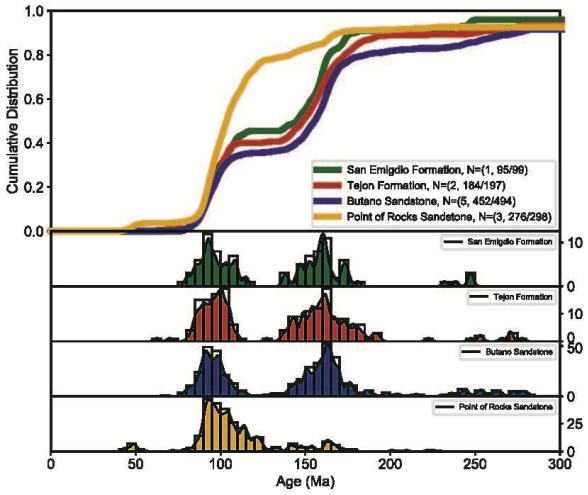
Sharman et al. - Figure 4



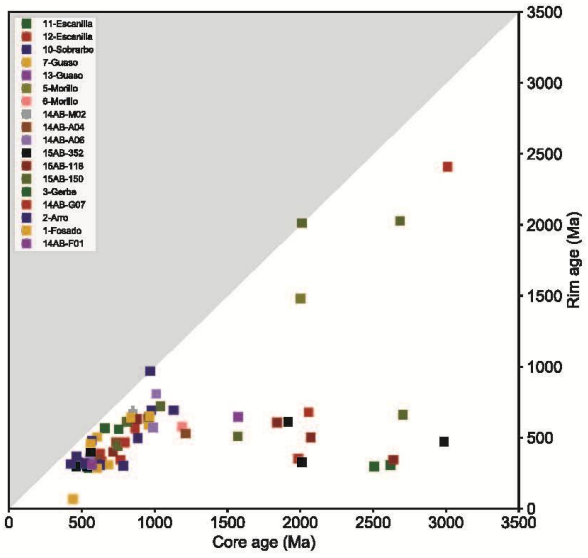
Sharman et al. - Figure 5



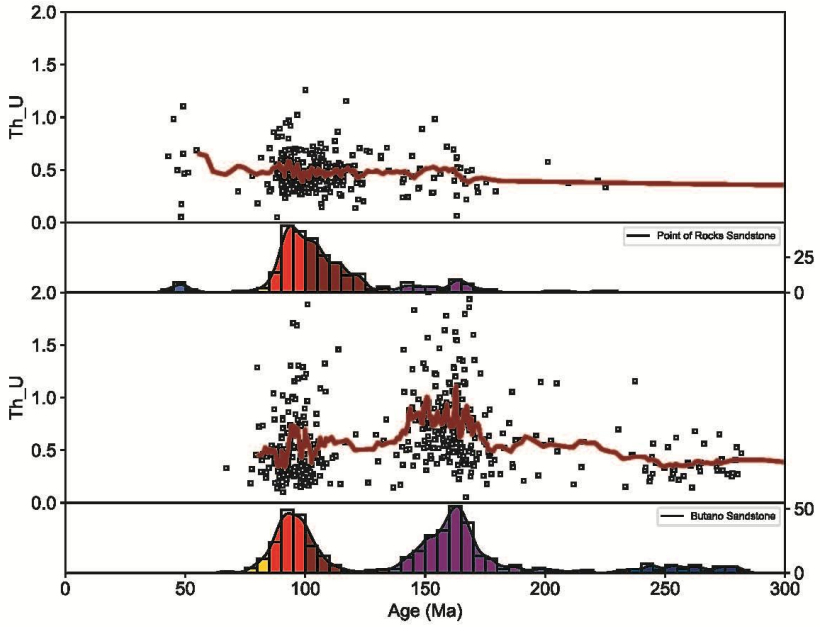
Sharman et al. - Figure 6



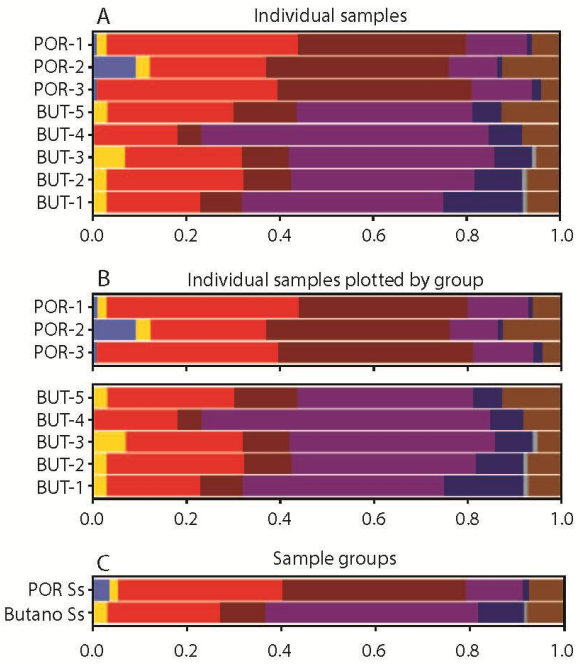
Sharman et al. - Figure 7



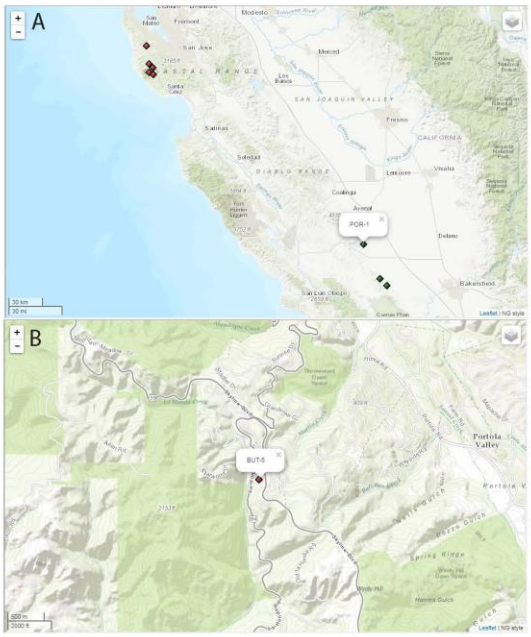
Sharman et al. - Figure 8



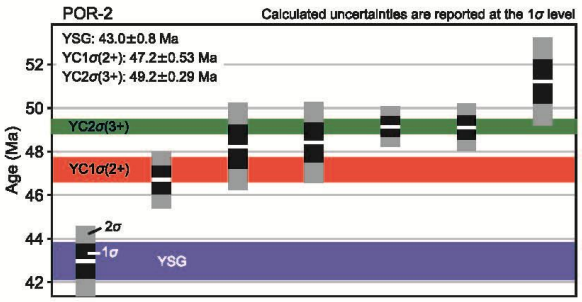
Sharman et al. - Figure 9



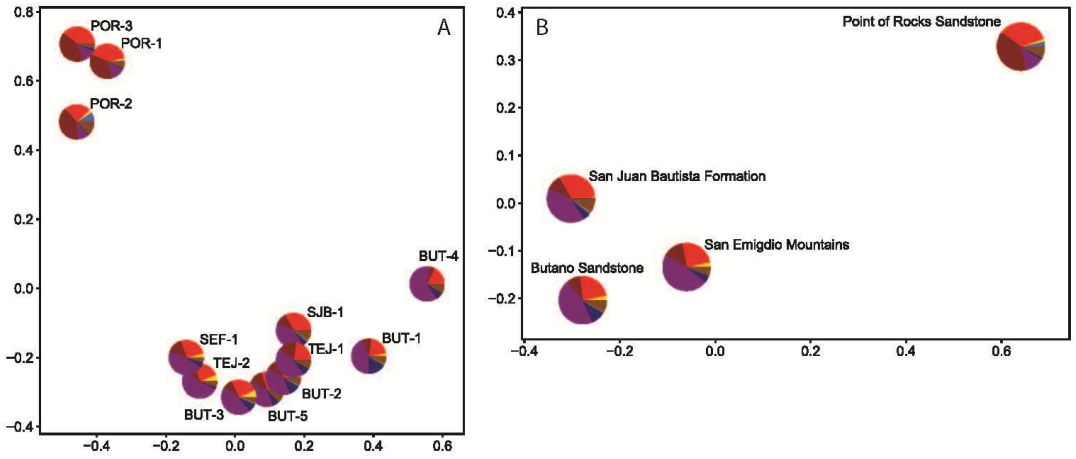
Sharman et al. - Figure 10



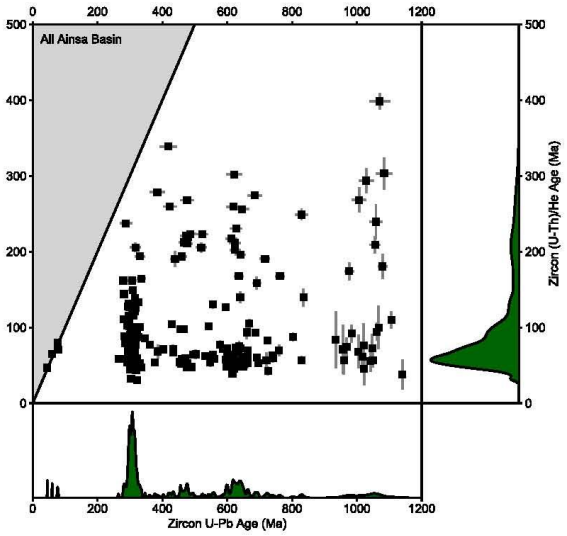
Sharman et al. - Figure 11



Sharman et al. - Figure 12



Sharman et al. - Figure 13



Sharman et al. - Figure 14