

RESEARCH ARTICLE

HTSSIP: An R package for analysis of high throughput sequencing data from nucleic acid stable isotope probing (SIP) experiments

Nicholas D. Youngblut^{1*}, Samuel E. Barnett, Daniel H. Buckley

School of Integrative Plant Science, Cornell University, Ithaca, New York, United States of America

¹ Current address: Department of Microbiome Science, Max Planck Institute for Developmental Biology, Tübingen, Germany

* nyoungblut@tuebingen.mpg.de



OPEN ACCESS

Citation: Youngblut ND, Barnett SE, Buckley DH (2018) HTSSIP: An R package for analysis of high throughput sequencing data from nucleic acid stable isotope probing (SIP) experiments. PLoS ONE 13(1): e0189616. <https://doi.org/10.1371/journal.pone.0189616>

Editor: Chongle Pan, Oak Ridge National Laboratory, UNITED STATES

Received: July 26, 2017

Accepted: November 29, 2017

Published: January 3, 2018

Copyright: © 2018 Youngblut et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are available within the R package that this paper describes. The R package (plus the associated data) is publicly available at two locations: CRAN (<https://cran.r-project.org/web/packages/HTSSIP/index.html>) and GitHub (<https://github.com/buckleylab/HTSSIP>).

Funding: This work is supported by the Department of Energy Office of Science, Office of Biological & Environmental Research Genomic Science Program (<http://genomicscience.energy>).

Abstract

Combining high throughput sequencing with stable isotope probing (HTS-SIP) is a powerful method for mapping *in situ* metabolic processes to thousands of microbial taxa. However, accurately mapping metabolic processes to taxa is complex and challenging. Multiple HTS-SIP data analysis methods have been developed, including high-resolution stable isotope probing (HR-SIP), multi-window high-resolution stable isotope probing (MW-HR-SIP), quantitative stable isotope probing (qSIP), and Δ BD. Currently, there is no publicly available software designed specifically for analyzing HTS-SIP data. To address this shortfall, we have developed the *HTSSIP* R package, an open-source, cross-platform toolset for conducting HTS-SIP analyses in a straightforward and easily reproducible manner. The *HTSSIP* package, along with full documentation and examples, is available from CRAN at <https://cran.r-project.org/web/packages/HTSSIP/index.html> and Github at <https://github.com/buckleylab/HTSSIP>.

Introduction

Stable isotope probing of nucleic acids (DNA- and RNA-SIP) is a powerful method for mapping *in situ* metabolic processes, such as nitrogen and carbon cycling, to microbial taxa. Historically, the sensitivity of nucleic acid SIP has been limited by the low throughput of DNA sequencing and the low taxonomic resolution of DNA fingerprinting techniques [1,2]. Recently, DNA- and RNA-SIP have been combined with high throughput sequencing of PCR amplicons (HTS-SIP), which allows researchers to map *in situ* metabolic processes to thousands of taxa resolved at a fine taxonomic resolution [3–5].

While HTS-SIP is proving to be a very useful method for exploring *in situ* metabolic processes in complex microbial communities, the accurate analysis of HTS-SIP datasets is complex [6,7]. Multiple strategies have been developed for analyzing HTS-SIP data, including high-resolution stable isotope probing (HR-SIP) [5], multi-window high-resolution stable isotope probing (MW-HR-SIP) [7], quantitative stable isotope probing (qSIP) [3], and Δ BD (delta Buoyant Density) [5]. Prior to the development of these HTS-SIP analysis methods,

gov/) under award numbers DE-SC0004486 and DE-SC0010558, with Dr. Daniel Buckley being the recipient of both awards. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

DNA- and RNA-SIP experiments that utilized Sanger or high throughput sequencing were usually analyzed with standard statistical processes (e.g. t-tests), which have low sensitivity and/or high false positive rates when applied to sequence data [7,8]. The goals of HTS-SIP analysis methods differ, with HR-SIP and MW-HR-SIP designed to accurately identify taxa that have incorporated isotopically labeled substrate (i.e. ‘incorporators’), while the main goal of qSIP and Δ BD is to quantify the amount of isotopic enrichment for each taxon (i.e. atom % excess). While all methods use amplicon sequences (e.g. 16S rRNA or fungal ITS sequences) from multiple fractions of each isopycnic gradient, HR-SIP, MW-HR-SIP, and Δ BD solely use sequence data while qSIP additionally requires qPCR derived estimations of gene copy number from each gradient fraction. Recently, Youngblut and Buckley developed a HTS-SIP simulation model and showed that MW-HR-SIP is more accurate for identifying incorporators than HR-SIP and qSIP, while qSIP is generally more precise than Δ BD for quantifying isotopic enrichment [7].

Currently, there are no publicly available software tools specifically designed for HTS-SIP analysis, which forces researchers to either recreate the analysis methods described in existing HTS-SIP studies or develop ad-hoc analysis methods. Both of these options decrease the ease of use and reproducibility of HTS-SIP analyses. To address this deficiency, we developed the *HTSSIP* R package, which includes the following features that aid in HTS-SIP analyses:

- Functions for conducting HR-SIP, MW-HR-SIP, qSIP, and Δ BD to analyze data from DNA-SIP and RNA-SIP experiments, plus functions for conducting historically used statistical methods (e.g. t-tests comparing taxon abundance in “heavy” fractions).
- Functions for performing HTS-SIP dataset simulation, as described [7]
- Functions for exploratory analysis of simulated HTS-SIP data, useful for predicting how different experimental designs can alter experimental outcomes
- Functions for exploratory analysis of real HTS-SIP data, useful for conducting post-hoc analyses
- Ability to run analyses with parallel processing
- Extensive documentation and tutorials (see the *HTSSIP* package vignettes)

Package description

Input data

Dataset input is handled by the *Phyloseq* R package, a feature-rich package for general microbiome data analysis that can be used to import many common microbiome data formats [9]. *HTSSIP* includes convenience functions to easily and flexibly designate the experimental design of the SIP experiment for downstream HTS-SIP analyses (Fig 1).

HTS-SIP dataset exploratory analyses

A common first step in analyzing nucleic acid SIP data is to quantify the total nucleic acid concentration or gene copy number (estimated by qPCR) across density gradients in order to determine the buoyant density (BD) “shift” of nucleic acids in isotopically labeled treatments versus unlabeled controls [10,11]. The general expectation is that a “shift” of nucleic acid BD from “light” towards “heavy” densities is indicative of isotope incorporation. However, in a well designed SIP experiment, the ratio of added to endogenous substrate should be small, and

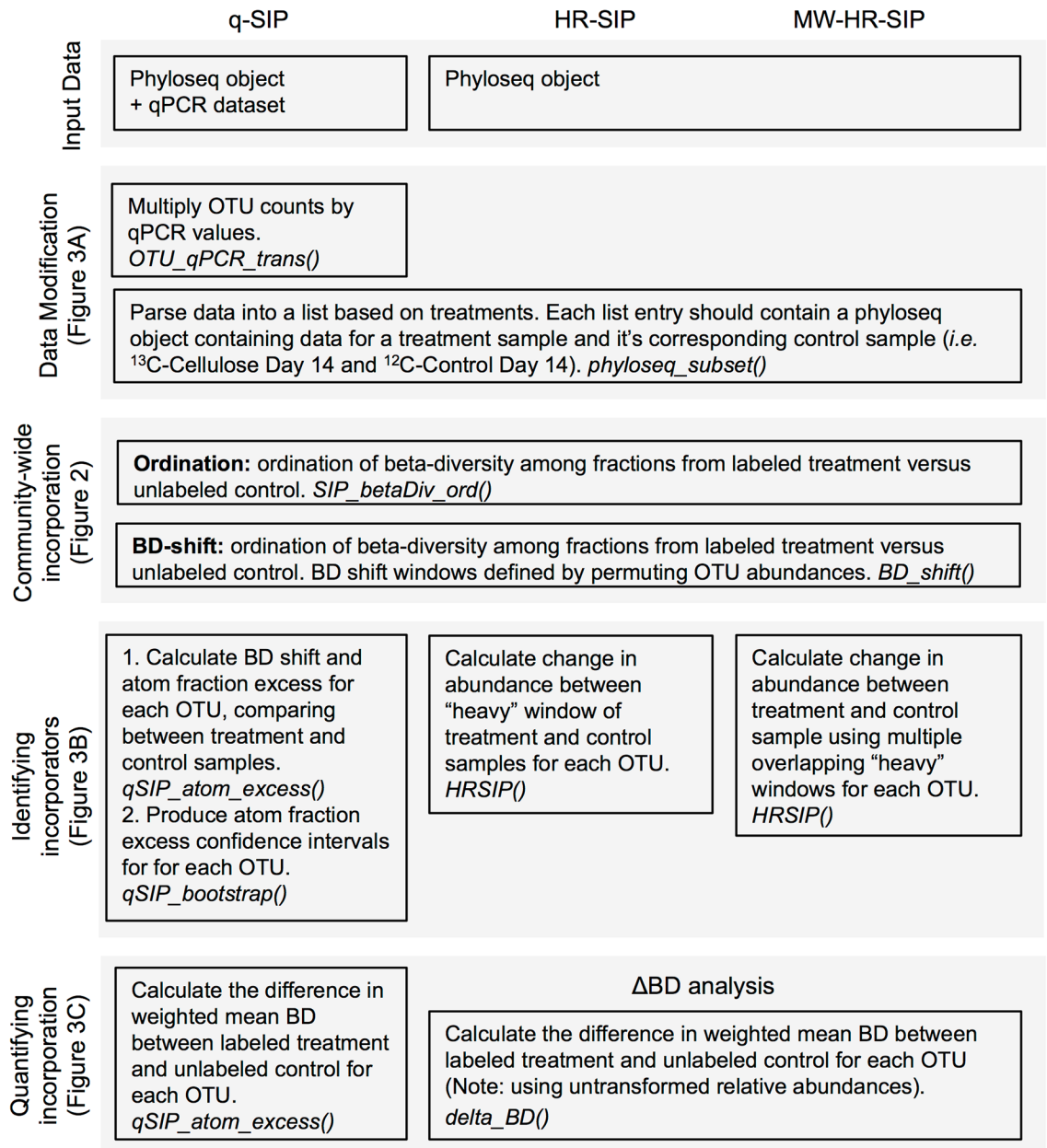


Fig 1. A diagram depicting the possible HTS-SIP analyses available in the HTSSIP R package. The R functions to conduct each workflow step are italicized, and the figure references refer to example data produced by these workflow steps.

<https://doi.org/10.1371/journal.pone.0189616.g001>

this can produce an imperceptible BD shift [4]. In addition, an extensive shift may indicate excessive cross-feeding [12]. HTS-SIP methods can detect taxa that have incorporated low levels of isotope, or occur at frequencies that are so low that they do not cause a shift in the overall BD of community nucleic acids [5]. As a result, analysis of the BD distribution of total nucleic acids within density gradients is of little utility in assessing the results of nucleic acid SIP experiments performed on complex communities.

As a simpler alternative, which leverages the power of high-throughput sequencing techniques, BD “shifts” can be inferred solely from sequence data [4,5]. Given that incorporators

will be more abundant in “heavy” gradient fractions of the labeled treatment versus the unlabeled control, a BD shift can be inferred by assessing the beta-diversity between treatment and control gradient fractions. This approach is more sensitive for detecting community-level isotope incorporation than the approach of quantifying total nucleic acid concentration across the density gradient [7]. The *HTSSIP* R package implements two methods for using beta-diversity to assess isotope incorporation at the community-level: an ordination approach and an approach that expresses beta-diversity between corresponding treatment and control fractions as a function of their BD (Fig 1).

The ordination approach simply involves pairwise calculations of a beta-diversity metric between all gradient fractions from isotopically labeled treatments and corresponding unlabeled controls, followed by visualizing the distance matrix with either principal coordinates analysis (PCoA) or non-metric multidimensional scaling (NMDS). An increase in beta-diversity between corresponding gradient fractions of labeled samples and controls is expected if isotope incorporation causes a change in the BD of OTUs (Fig 2A & 2B).

While the ordination approach provides a useful overview of community-wide isotope incorporation, the extent of incorporation is difficult to compare among multiple treatments (e.g. ^{13}C -cellulose versus ^{13}C -xylose). The second approach implemented in the *HTSSIP* package visualizes DNA BD shifts by calculating pairwise beta-diversity of corresponding gradient fractions between treatment and control gradients (or control versus control to assess technical variation among replicates). To deal with partially overlapping gradient fractions between gradients, the weighted mean beta-diversity is calculated from all treatment gradient fractions that overlap each control gradient fraction, with weights defined as the percent overlap in the BD range of each fraction (Fig 2C & 2D). A permutation test is used to identify BD ranges of high beta-diversity resulting from BD shifts (“BD shift windows”). The permutation test involves constructing bootstrap confidence intervals (CI) of beta-diversity by permuting OTU abundances among control fractions to generate a null model “treatment” that is compared against the actual control fractions. This null model represents the baseline beta-diversity expected if the treatment was just a randomly shuffled version of the control. A note in interpreting these data is that isotope incorporation will cause DNA to shift out of “light” gradient fractions and into “heavy” gradient fractions. Hence, in the presence of isotope incorporation, high beta-diversity can be observed in both “heavy” and “light” gradient fractions. Alternatively, in the absence of isotope incorporation, beta-diversity will remain low across all gradient fractions.

Identifying incorporators

HR-SIP, MW-HR-SIP, and qSIP can all be used to identify incorporators. To illustrate the application of HR-SIP, MW-HR-SIP, and qSIP in the *HTSSIP* R package, we simulated a simplified HTS-SIP dataset consisting of 10 OTUs (Fig 3A). Our purpose here is merely to illustrate functions of the *HTSSIP* R package; comprehensive assessment of the accuracy of these techniques (and historically used DNA- and RNA-SIP analysis methods) is available elsewhere [7]. Briefly, HR-SIP identifies incorporators by utilizing DESeq2 to identify OTUs that have high differential relative abundance in “heavy” fractions of labeled treatment versus unlabeled control [8]. MW-HR-SIP takes the same relative abundance based approach as HR-SIP but uses multiple overlapping “heavy” BD windows (while correcting for multiple hypotheses). In contrast, qSIP uses qPCR data to transform OTU relative abundance distributions into pseudo-absolute abundance distributions (Fig 3A), and then BD shifts are determined from these transformed distributions by calculating the difference in center of mass for each OTU in treatment versus control gradients. Atom fraction excess can thus be calculated for specific

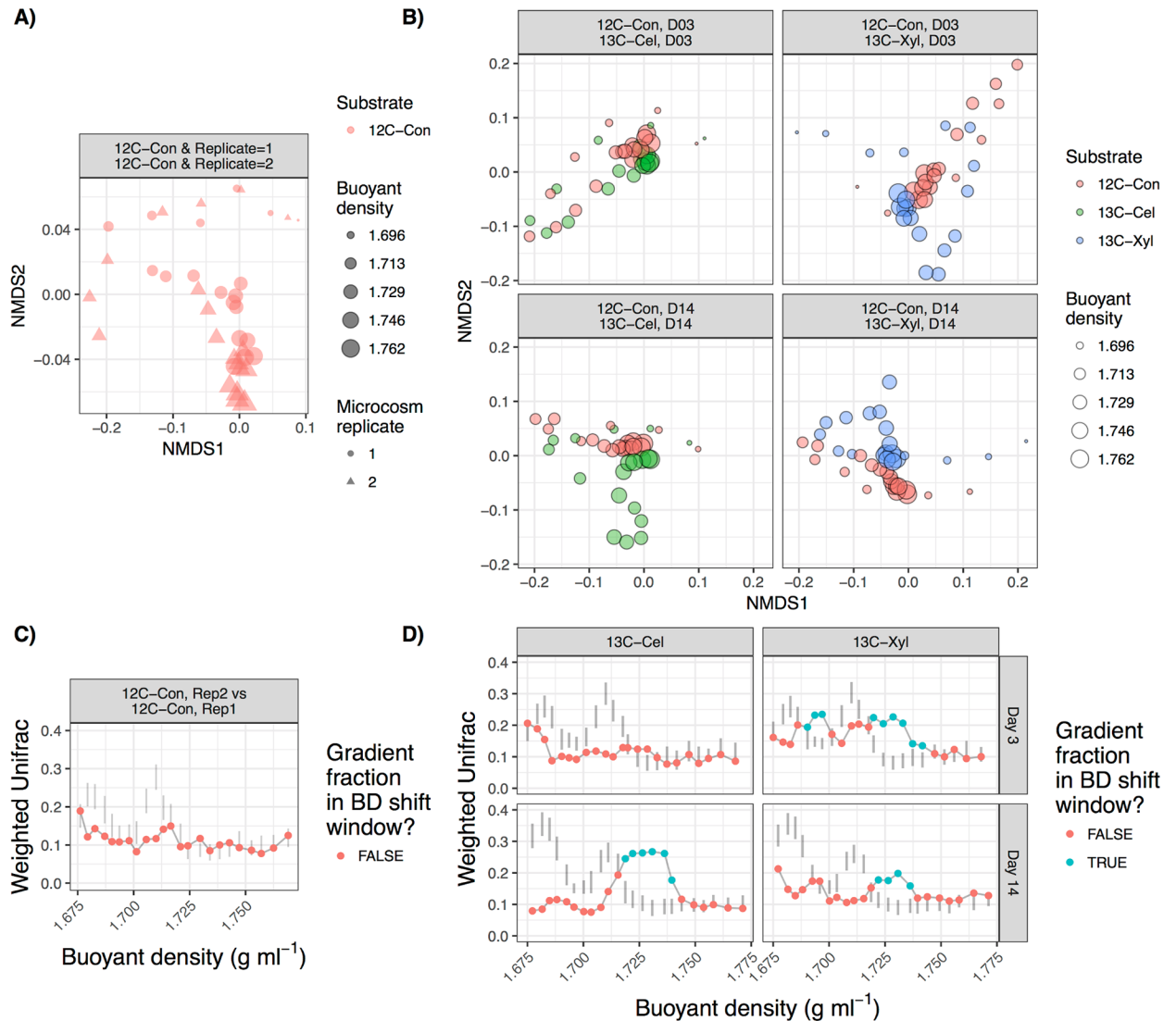


Fig 2. Examples of analyses available in the HTSSIP R package for visualizing results from HTS-SIP experiments. These figures were generated using data from an HTS-SIP experiment that consisted of 3 treatments where only the isotopic label was varied: an unlabeled control (12C-Con), ¹³C-xylose (13C-Xyl), and ¹³C-cellulose (13C-Cel) [7]. DNA was extracted from each treatment 3 and 14 days after substrate addition, subjected to CsCl fractionation, and 16S rRNA sequencing was performed on ~24 fractions per CsCl gradient. Analyzing beta-diversity between gradient fractions can reveal changes in the buoyant density (BD) distribution of DNA in ¹³C-labeled treatments versus the unlabeled control. Panel A is a non-metric multidimensional scaling (NMDS) ordination comparing beta-diversity (weighted Unifrac) between gradient fractions from two unlabeled replicates, which indicates that sequence composition varies greatly with BD but not between experimental replicates. Panel B is an NMDS ordination similar to Panel A, but each facet contains a ¹³C treatment and the corresponding unlabeled control. Large relative distance between “heavy” control and labeled gradient fractions indicates a change in community composition caused by shift in DNA BD that resulted from ¹³C isotope incorporation. The NMDS stress values ranged from 0.06 to 0.07. Panels C and D depict the same data as in Panels A and B, respectively, but only depict beta-diversity among gradient fractions that correspond in BD between labeled treatment and unlabeled control fractions. In this visualization, increases in beta-diversity indicate a change in community composition caused by a shift in DNA BD resulting from ¹³C isotope incorporation. The line ranges represent 95% confidence intervals (CI) calculated by permuting OTU abundances and recalculating beta-diversity (100 bootstrap replicates), and actual beta-diversity values are represented by colored circles. “BD shift windows” (blue circles) indicate regions defined by ≥3 consecutive fractions with beta-diversities greater than expected by chance (*i.e.* exceeding the CI).

<https://doi.org/10.1371/journal.pone.0189616.g002>

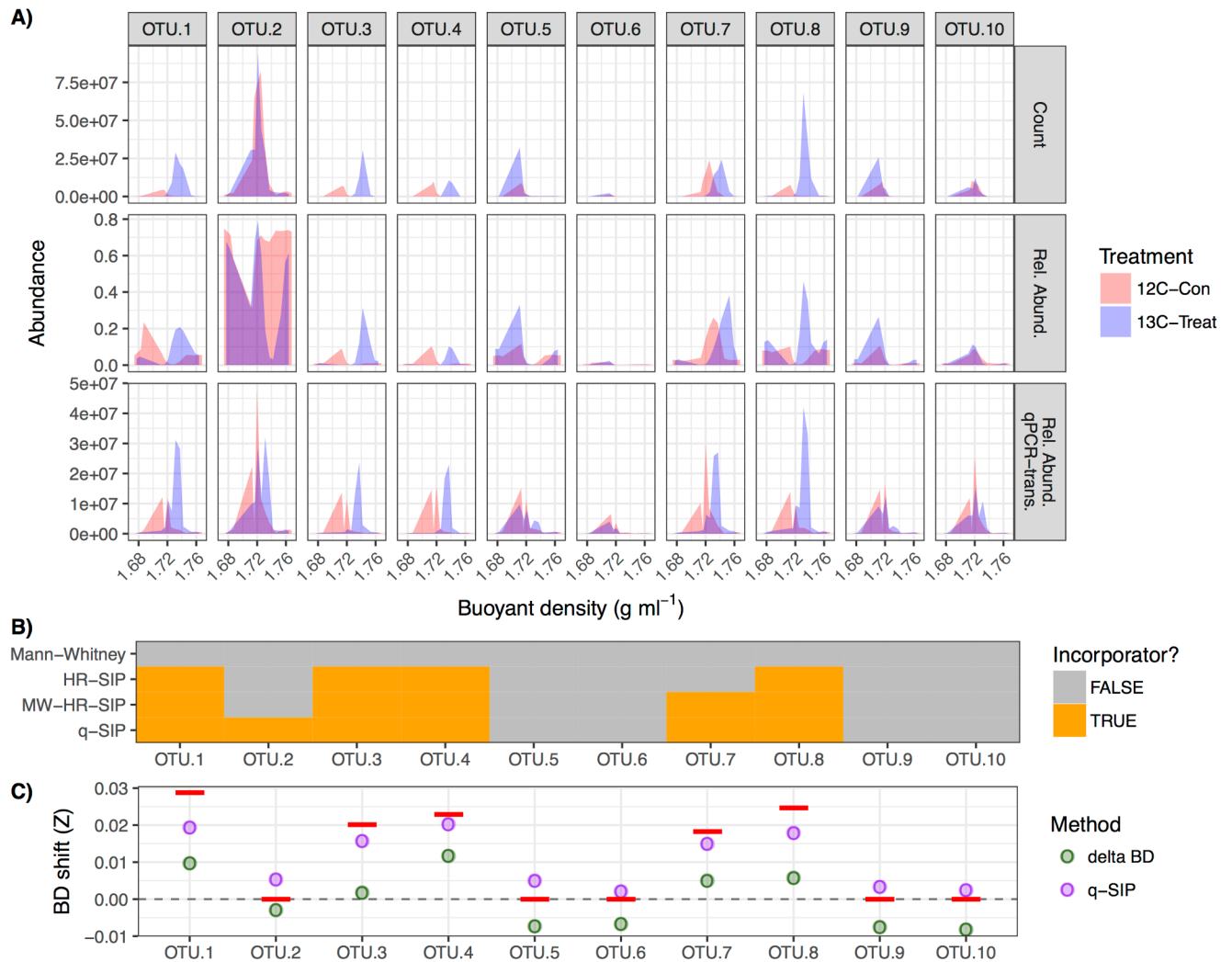


Fig 3. Examples of using the HTSSIP R package for data processing, data exploration, incorporator identification, and quantification of BD shifts. The SIPsim toolset was used to simulate a simple HTSSIP dataset consisting of two treatments: a ¹³C-treatment (“13C-Treat”) and a ¹²C-control (“12C-Con”). Each treatment has 3 replicates, with each consisting of 24 gradient fractions [7]. Half of the ten OTUs (OTU 1, 3, 4, 7, and 8) were randomly assigned an atom excess ¹³C between 30 and 100%. [3]. Panel A depicts the raw abundances (“Counts”), fractional relative abundance (“Rel. Abund.”), and relative abundances transformed by simulated qPCR data (“Rel. Abund. qPCR-trans.”). For clarity, only 1 of the 3 experimental replicates is shown. Panel B shows those OTUs identified as “incorporators” (incorporated ¹³C into genomic DNA) by HR-SIP, MW-HR-SIP, qSIP, or by using a “heavy-SIP” approach (the Mann Whitney U test). Panel C shows the BD shift of each OTU as determined by ΔBD or qSIP. The dashed line signifies a BD shift (Z) of 0.0 g ml⁻¹, and the red bars show the true theoretical BD shift resulting from ¹³C isotope incorporation.

<https://doi.org/10.1371/journal.pone.0189616.g003>

isotopes (e.g. ¹³C or ¹⁵N) based on the calculations described in the work of Hungate and colleagues [3]. In order to identify incorporators, a permutation test is used to construct bootstrap confidence intervals of atom fraction excess. The HTSSIP package also provides functions for conducting historically used DNA- and RNA-SIP analysis methods (e.g. using t-tests to compare OTU abundances between “heavy” control versus treatment fractions). We refer to these methods collectively as “heavy-SIP”, which have been extensively compared to HR-SIP methods in other work [7]. Sensitivity in identifying incorporators can depend on the methods used (Fig 3B), and although we only provide a simple comparison of methods here, the sensitivity and specificity of these results reflect those of Youngblut and Buckley [7]. SIP experiments

can be simulated using the SIPSIm toolset [7], and these data analyzed using the *HTS-SIP* R package. Such *in silico* evaluation is valuable for predicting possible experimental outcomes and the expected analytical accuracy of SIP experiments based on details of experimental design prior to conducting experiments.

The *HTSSIP* package implements HR-SIP based on the code provided in the work of Pepe-Ranney and colleagues [5]. MW-HR-SIP is implemented in *HTSSIP* based on the R code provided in the SIPSIm HTS-SIP dataset simulation toolset [7]. The *HTSSIP* implementation of qSIP is based on the method's description in the work of Hungate and colleagues [3]. Implementations of each method include the option for parallel processing of each algorithm. Parallelization is implemented through the *plyr* R package [13], which allows for various parallel backends to be used such as *doSNOW* and *doParallel*.

Quantifying isotopic enrichment

Unlike HR-SIP and MW-HR-SIP, the main goal of qSIP and Δ BD is to quantify isotopic enrichment. To illustrate the use of the *HTSSIP* package for conducting qSIP and Δ BD, we applied both analyses to the simplified HTS-SIP dataset described above (Fig 3C). Δ BD is implemented in *HTSSIP* as described in the work of Pepe-Ranney and colleagues [5]. As shown in Youngblut and Buckley [7], qSIP and Δ BD can produce substantially different estimates of isotope incorporation.

Simulating datasets

The *HTSSIP* package provides functions to simulate simple HTS-SIP datasets for use in software testing, analysis pipeline development, and gaining familiarity with software and data formats. However, the SIPSIm toolset is recommended for evaluating possible SIP experimental designs and for testing the accuracy of HTS-SIP analyses, because the simulation framework for SIPSIm is based on the physics of isopycnic centrifugation, unlike the simulations possible with *HTSSIP* [7]. *HTSSIP* utilizes *coenocliner*, an R package designed for simulating taxon abundance across environmental gradients, to simulate taxon abundance distributions across buoyant density gradient fractions [13].

Future work

Future development of the *HTSSIP* package will include i) functions for mapping incorporator status to phylogenies and visualizing the results ii) direct integration with the SIPSIm toolset for rapid HTS-SIP experimental design and assessment of accuracy iii) functions analyzing shotgun metagenome data derived from SIP experiments.

Conclusions

Given the power of HTS-SIP for mapping *in situ* metabolism to taxonomic identity, adoption of the technique by researchers will greatly help to resolve connections between microbial ecology and taxonomy. Currently, HTS-SIP data analysis is complex, with few existing computational tools to aid researchers. The R package *HTSSIP* provides a single, standardized analysis pipeline that facilitates reproducible analyses on HTS-SIP datasets and direct cross-study comparisons. Moreover, *HTSSIP* can be combined with the SIPSIm toolset to simulate and evaluate possible DNA-SIP experimental designs.

Author Contributions

Conceptualization: Nicholas D. Youngblut, Daniel H. Buckley.

Data curation: Nicholas D. Youngblut.

Formal analysis: Nicholas D. Youngblut.

Funding acquisition: Daniel H. Buckley.

Investigation: Nicholas D. Youngblut.

Methodology: Nicholas D. Youngblut, Daniel H. Buckley.

Project administration: Daniel H. Buckley.

Resources: Samuel E. Barnett, Daniel H. Buckley.

Software: Nicholas D. Youngblut, Samuel E. Barnett.

Supervision: Daniel H. Buckley.

Visualization: Nicholas D. Youngblut.

Writing – original draft: Nicholas D. Youngblut, Daniel H. Buckley.

Writing – review & editing: Nicholas D. Youngblut, Samuel E. Barnett, Daniel H. Buckley.

References

1. Whiteley AS, Thomson B, Lueders T, Manefield M. RNA stable-isotope probing. *Nat Protoc.* 2007; 2: 838–844. <https://doi.org/10.1038/nprot.2007.115> PMID: 17446884
2. Uhlík O, Jecná K, Leigh MB, Macková M, Macek T. DNA-based stable isotope probing: a link between community structure and function. *Sci Total Environ.* 2009; 407: 3611–3619. <https://doi.org/10.1016/j.scitotenv.2008.05.012> PMID: 18573518
3. Hungate BA, Mau RL, Schwartz E, Caporaso JG, Dijkstra P, van Gestel N, et al. Quantitative microbial ecology through stable isotope probing. *Appl Environ Microbiol.* 2015; 81: 7570–7581. <https://doi.org/10.1128/AEM.02280-15> PMID: 26296731
4. Pepe-Ranney C, Koechli C, Potrafka R, Andam C, Eggleston E, Garcia-Pichel F, et al. Non-cyanobacterial diazotrophs mediate dinitrogen fixation in biological soil crusts during early crust formation. *ISME J.* 2016; 10: 287–298. <https://doi.org/10.1038/ismej.2015.106> PMID: 26114889
5. Pepe-Ranney C, Campbell AN, Koechli CN, Berthrong S, Buckley DH. Unearthing the ecology of soil microorganisms using a high resolution DNA-SIP approach to explore cellulose and xylose metabolism in soil. *Front Microbiol.* 2016; 7: 703. <https://doi.org/10.3389/fmicb.2016.00703> PMID: 27242725
6. Youngblut ND, Buckley DH. Intra-genomic variation in G + C content and its implications for DNA stable isotope probing. *Environ Microbiol Rep.* 2014; 6: 767–775. <https://doi.org/10.1111/1758-2229.12201> PMID: 25139123
7. Youngblut ND, Buckley DH. Evaluating the accuracy of DNA stable isotope probing. *bioRxiv.* 2017. p. 138719. <https://doi.org/10.1101/138719>
8. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014; 15: 550. <https://doi.org/10.1186/s13059-014-0550-8> PMID: 25516281
9. McMurdie PJ, Holmes S. phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE.* 2013. p. e61217. Available: <http://dx.plos.org/10.1371/journal.pone.0061217> PMID: 23630581
10. Lueders T, Kindler R, Miltner A, Friedrich MW, Kaestner M. Identification of bacterial micropredators distinctively active in a soil microbial food web. *Appl Environ Microbiol.* 2006; 72: 5342–5348. <https://doi.org/10.1128/AEM.00400-06> PMID: 16885285
11. El Zahar Haichar F, Achouak W, Christen R, Heulin T, Marol C, Marais M-F, et al. Identification of cellulolytic bacteria in soil by stable isotope probing. *Environ Microbiol.* 2007; 9: 625–634. <https://doi.org/10.1111/j.1462-2920.2006.01182.x> PMID: 17298363
12. DeRito CM, Pumphrey GM, Madsen EL. Use of field-based stable isotope probing to identify adapted populations and track carbon flow through a phenol-degrading soil microbial community. *Appl Environ Microbiol.* 2005; 71: 7858–7865. <https://doi.org/10.1128/AEM.71.12.7858-7865.2005> PMID: 16332760
13. Wickham H. The split-apply-combine strategy for data analysis. *Journal of Statistical Software, Articles.* 2011; 40: 1–29.