Bioinformatics, 2024, **40(7)**, btae448
https://doi.org/10.1093/bioinformatics/btae448
Advance Access Publication Date: 11 July 2024
Original Paper

Sequence analysis

isolateR: an R package for generating microbial libraries from Sanger sequencing data

Brendan Daisley[1,†,*], Sarah J. Vancuren[1,†], Dylan J.L. Brettingham[1], Jacob Wilde[1], Simone Renwick[2,3], Christine V. Macpherson[1], David A. Good[1], Alexander J. Botschner[1], Sandi Yen[4], Janet E. Hill[5], Matthew T. Sorbara[1], Emma Allen-Vercoe[1]

[1]Department of Molecular and Cellular Biology, University of Guelph, Guelph, ON N1G 2W1, Canada
[2]Department of Pediatrics, School of Medicine, University of California, San Diego, United States
[3]Larsson-Rosenquist Foundation Mother-Milk-Infant Center of Research Excellence (MOMI CORE), The Human Milk Institute (HMI), University of California, San Diego, CA 92093, United States
[4]Kennedy Institute of Rheumatology, Medical Sciences Division, University of Oxford, Oxford OX1 2JD, United Kingdom
[5]Department of Veterinary Microbiology, University of Saskatchewan, Saskatoon, SK S7N 5B4, Canada

*Corresponding author. Department of Molecular and Cellular Biology, University of Guelph, Guelph, ON N1G 2W1, Canada. E-mail: bdaisley@uoguelph.ca (B.D.)

[‡]= equal contribution.

Associate Editor: Russell Schwartz

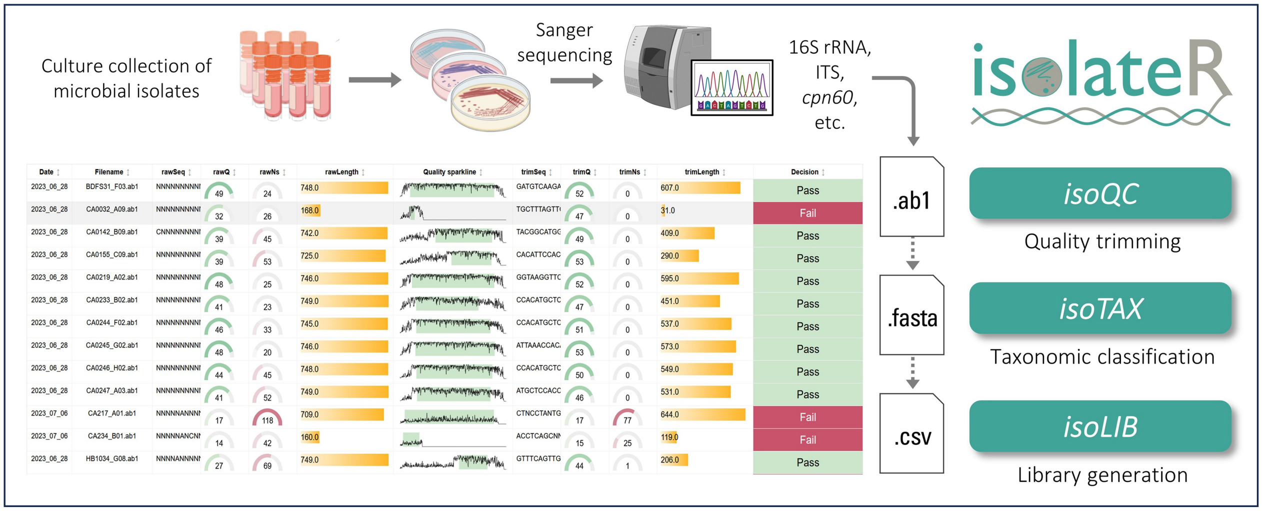
Abstract

Motivation: Sanger sequencing of taxonomic marker genes (e.g. 16S/18S/ITS/*rpoB*/*cpn60*) represents the leading method for identifying a wide range of microorganisms including bacteria, archaea, and fungi. However, the manual processing of sequence data and limitations associated with conventional BLAST searches impede the efficient generation of strain libraries essential for cataloging microbial diversity and discovering novel species.

Results: isolateR addresses these challenges by implementing a standardized and scalable three-step pipeline that includes: (1) automated batch processing of Sanger sequence files, (2) taxonomic classification *via* global alignment to type strain databases in accordance with the latest international nomenclature standards, and (3) straightforward creation of strain libraries and handling of clonal isolates, with the ability to set customizable sequence dereplication thresholds and combine data from multiple sequencing runs into a single library. The tool's user-friendly design also features interactive HTML outputs that simplify data exploration and analysis. Additionally, *in silico* benchmarking done on two comprehensive human gut genome catalogues (IMGG and Hadza hunter-gather populations) showcase the proficiency of isolateR in uncovering and cataloging the nuanced spectrum of microbial diversity, advocating for a more targeted and granular exploration within individual hosts to achieve the highest strain-level resolution possible when generating culture collections.

Availability and implementation: isolateR is available at: https://github.com/bdaisley/isolateR.

Graphical abstract



Date	Filename	rawSeq	rawQ	rawNs	rawLength	Quality sparkline	trimSeq	trimQ	trimNs	trimLength	Decision
2023_06_28	BDFS31_F03.ab1	NNNNNNNNNN	49	24	748.0		GATGTCAAGA	52	0	607.0	Pass
2023_06_28	CA0032_A09.ab1	NNNNNNNNNN	32	26	168.0		TGCTTTAGTT	47	0	31.0	Fail
2023_06_28	CA0142_B09.ab1	CNNNNNNNNN	39	45	742.0		TACGGCATGG	49	0	409.0	Pass
2023_06_28	CA0155_C09.ab1	NNNNNNNNNN	39	53	725.0		CACATTCCAC	53	0	290.0	Pass
2023_06_28	CA0219_A02.ab1	NNNNNNNNNN	48	25	746.0		GGTAAGGTTC	52	0	595.0	Pass
2023_06_28	CA0233_B02.ab1	NNNNNNNNNN	41	23	749.0		CCACATGCTC	47	0	451.0	Pass
2023_06_28	CA0244_F02.ab1	NNNNNNNNNN	46	33	745.0		CCACATGCTC	51	0	537.0	Pass
2023_06_28	CA0245_G02.ab1	NNNNNNNNNN	48	20	746.0		ATTAAACCAC	53	0	573.0	Pass
2023_06_28	CA0246_H02.ab1	NNNNNNNNNN	44	45	748.0		CCACATGCTC	50	0	549.0	Pass
2023_06_28	CA0247_A03.ab1	NNNNNNNNNN	41	52	749.0		ATGCTCCACC	46	0	531.0	Pass
2023_07_06	CA217_A01.ab1	NNNNNANNNN	17	118	709.0		CTNCCTANTG	17	77	644.0	Fail
2023_07_06	CA234_B01.ab1	NNNNNANCNN	14	42	160.0		ACCTCAGCNN	15	25	119.0	Fail
2023_06_28	HB1034_G08.ab1	NNNNANNNNN	27	69	749.0		GTTTCAGTTG	44	1	206.0	Pass

Received: 30 April 2024; **Revised:** 6 July 2024; **Editorial Decision:** 9 July 2024; **Accepted:** 10 July 2024

1 Introduction

The rapid expansion of culturomics in microbial research has emphasized the need for standardized generation of comprehensive strain libraries to catalogue the extensive variety of microorganisms identified within different environments. A single culture-based investigation can typically yield over a thousand different microbial isolates (Lagier *et al.* 2016), underscoring the challenge of processing and analyzing the vast amount of associated sequencing data. As the gold standard in microbial genomics, Sanger sequencing (Sanger *et al.* 1977) produces high-accuracy sequence data stored in ABIF formatted trace files. Several software packages including Poly Peak Parser (Hill *et al.* 2014), sangeranalyseR (Chao *et al.* 2021), and TraceTrack (Brazdilova *et al.* 2023) have recently been developed to improve aspects of analysis such as basecalling, visualization of chromatograms, and batch processing of data. Despite these advancements, there is a notable lack of software available to organize integrated taxonomic classification of Sanger sequence data or the standardized generation of strain libraries, both of which are crucial aspects limiting current microbial isolation workflows.

The 16S rRNA gene is the most commonly used marker for the taxonomic classification of cultured and uncultured bacteria and archaea (Janda and Abbott 2007). However, other marker genes such as *rpoB* (Case *et al.* 2007) and *cpn60* (Vancuren and Hill 2019) have emerged as valuable tools for their ability to delineate taxonomy with greater resolution. Additionally, for fungal isolation and classification, 18S rRNA and Internal Transcribed Spacer (ITS) regions serve as critical taxonomic markers, offering insights into fungal diversity and phylogeny (Schoch *et al.* 2012).

In terms of assigning taxonomy to marker sequence data, conventional methods typically rely on the Basic Local Alignment Search Tool (BLAST) and GenBank non-redundant nucleotide databases which contain sequences from non-type strain material (Altschul *et al.* 1997). While this approach generally offers a satisfactory estimate of taxonomy, relying on comparisons to non-type strain material, such as uncultured clone sequences, can yield high-confidence matches that overlook sequences indicative of novel species. Moreover, local alignment algorithms (e.g. the seed-and-extend heuristic used in BLAST) have limitations in accurately classifying distantly-related sequences (Tindall *et al.* 2010). Global alignment algorithms typically provide better phylogenetic discrimination. For example, clear taxonomic boundaries have been determined for bacteria and archaea based on global alignment of 16S rRNA gene sequence similarity at the species (98.7%), genus (94.5%), family (86.5%), order (82.0%), class (78.5%), and phylum (75.0%) levels (Yarza *et al.* 2014). Similar hierarchical approaches have been applied in developing rational thresholds for delimiting filamentous fungi based on ITS sequence similarity (Vu *et al.* 2019), as well as the "species hypothesis" system used in the UNITE database (Nilsson *et al.* 2019). While sequence identity thresholds are not universally applicable to all groups of taxa nor all marker sequence regions (Edgar 2018), they serve as a useful general indicator for screening of potentially novel taxa in culture collections.

There is a pressing need for an automated system to handle large volumes of Sanger sequencing data efficiently and accurately. Such a system should not only streamline sequence quality control and trimming but also incorporate a robust framework for up-to-date taxonomic assignment. Traditional methodologies fall short in this aspect, especially given the dynamic nature of microbial taxonomy and the necessity for ongoing updates against authoritative resources such as the List of Prokaryotic names with Standing in Nomenclature (LPSN) (Parte 2014).

In response, we have developed a dedicated R package, isolateR, to meet the high-throughput demands of modern microbial isolation workflows (Fig. 1). This tool not only automates the quality processing of ABIF files but also integrates a global alignment method to ensure accurate taxonomic classification based on type strain reference databases, as well as aids in the organized collection of microbial sequence data. By enhancing the precision and throughput of cataloging microbial strains, isolateR enables a more comprehensive and accurate exploration of microbial diversity, paving the way for new discoveries and applications in various scientific fields.

2 Materials and methods

2.1 Software description

isolateR is designed to run entirely within an R environment from either the command line interface or through the command console of RStudio. It is compatible with all major operating systems, including Windows, macOS, and Linux. The package is efficiently designed to offer a user-friendly experience with minimal dependencies and quick installation. Detailed documentation and usage examples are available on the package's GitHub repository (https://github.com/bdaisley/isolateR). On a standard laptop computer (8 processors @ 1.60 GHz), isolateR can process 100 Sanger sequencing files, including taxonomic annotation and strain library creation in less than 10 min using under 8 GB of RAM.

2.2 Data import and Sanger sequence quality processing

The first stage of analysis in the pipeline is performed *via* the "*isoQC*" function, which facilitates the batch input and processing of Sanger sequencing-derived ABIF files (.ab1 extension). Briefly, the DNA sequence trace information from each file is extracted and then the sangeranalyseR package is wrapped to perform basecalling with a signal ratio cutoff as previously described (Chao *et al.* 2021). The default value (0.33) ensures only signals greater than 1/3 of the maximum signal ratio are considered and that weaker signals are labeled as "N" (i.e. ambiguous). Subsequently, the per-base Phred quality scores are extracted and then sequences are trimmed using an adaptive approach based on the sliding window algorithm implemented in Trimmomatic (Bolger *et al.* 2014). By default, the sliding window size is set to 15 bases, whereas the quality score cutoff is set to 2/3 of the maximum Phred score per sequence.

Users can adjust the parameters for both the basecalling and trimming steps to allow removal of low-quality data which may interfere with downstream taxonomic classification and library creation steps. In cases of paired sequencing, consensus sequences from forward and reverse read pairs can be automatically generated *via* error-tolerant overlapping pattern recognition using the "sanger_consensus" function. Contig assembly from multiple (greater than two) sequence pairs is also supported *via* regex-friendly file grouping features and automatic detection of sequence directionality and overlap.

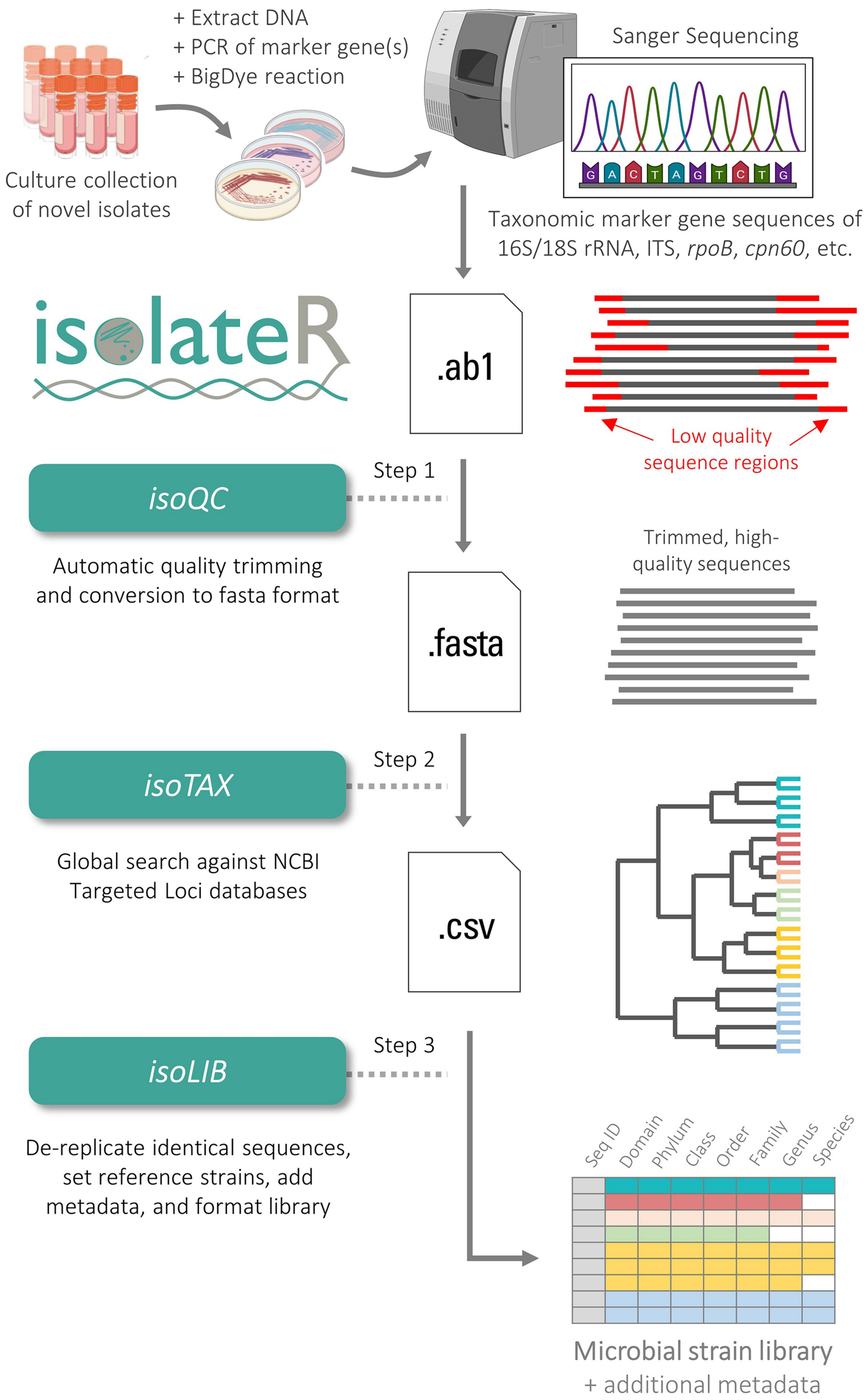

+ Extract DNA
+ PCR of marker gene(s)
+ BigDye reaction
Culture collection of novel isolates
Sanger Sequencing
G A C T A G T C T G
Taxonomic marker gene sequences of 16S/18S rRNA, ITS, rpoB, cpn60, etc.
isolateR
.ab1
Low quality sequence regions
Step 1
isoQC
Automatic quality trimming and conversion to fasta format
Trimmed, high-quality sequences
.fasta
Step 2
isoTAX
Global search against NCBI Targeted Loci databases
.csv
Step 3
isoLIB
De-replicate identical sequences, set reference strains, add metadata, and format library
Seq ID
Domain
Phylum
Class
Order
Family
Genus
Species
Microbial strain library
+ additional metadata

Figure 1. Overview of the isolateR command pipeline for microbial isolation workflows. The first step of the pipeline, *isoQC*, requires the input of taxonomic marker gene sequence trace files in ABIF format. The input sequences are automatically trimmed to remove poor quality reads (red sections) and then are categorized into PASS/FAIL groupings and converted to FASTA format for further inspection. The second step, *isoTAX*, performs taxonomic classification of the sequences by matching them against a type strain reference database for the marker gene of interest. Finally, the third step, *isoLIB*, determines the pairwise identity of sequences between one another to allow for standardized grouping of microbial isolates based on a desired similarity threshold. Ultimately, a microbial strain library is generated based on the dereplicated sequence representatives and then an interactive HTML table is output in a human-readable format. Information on the dereplicated isolates is retained. Quality metrics and other metadata about the sequences (including the dereplicated sequences) are retained and accessible through the menu buttons. Additional isolation runs can be added to the same strain library if desired and users can visualize phylogenetic relationships using several built-in functions of isolateR.

Sequences and their relevant metadata are then automatically output as an isoQC-class S4 object with slots containing the pre- and post-trimming sequence length, average quality scores, total number of N’s, and other metadata. A decision slot (pass/fail) is generated based on the final sequence quality characteristics; by default, those with a trimmed length of less than 200 bp and/or a mean Phred score of less than 20 are considered unacceptable quality. Final exports from this step include: (1) sequence files in FASTA format, (2) sequence metadata in CSV format, and (3) an interactive HTML table with descriptive statistics allowing inspection and fine tuning prior to moving to the next stage in the pipeline.

2.3 Taxonomic classification

The second stage of analysis is performed via the "*isoTAX*" function, which facilitates taxonomic classification based on pairwise sequence alignment and calculation of sequence similarity to the closest species representative from type strain material. This is achieved using an optimal global alignment approach, specifically by wrapping the Needleman–Wunsch algorithm implemented in VSEARCH (Rognes *et al.* 2016), in which pairwise similarity is defined as the total number of matching nucleotides in query and target sequences, excluding terminal gaps. By default, query sequences are assumed to represent a fragment of the 16S rRNA gene (i.e. the most common marker gene used for identifying bacterial or archaeal isolates) and are automatically matched against the latest version of the respective NCBI RefSeq Targeted Loci database, containing 16S rRNA sequences from bacteria and archaea type material only.

As an additional step to ensure compliance with international nomenclature standards, assigned genus and species names can also be screened against the LPSN database (Meier-Kolthoff *et al.* 2022), and automatically corrected if necessary.

Users can adjust several parameters to allow searching against other RefSeq Targeted Loci databases (e.g. ITS or 18S rRNA for identifying fungal isolates), as well as custom databases depending on the specific needs of an isolation workflow. In cases where a query sequence ambiguously matches with identical pairwise similarities to two or more database sequences representing different taxa, a decision is made to concatenate the species name and further inspection of the results are recommended. A list of closest matching species based on sequence identity are then output as an isoTAX-class S4 object with slots containing taxonomic information as well as other sequence metadata from the prior step (e.g. length, Phred quality, N's). Higher rank taxonomic information (phylum, class, order, and family) for each sequence is further retrieved using the NCBI Entrez batch search system (Winter 2017), and a unique column is designated for each rank from phylum to species level. Subsequently, the lowest common ancestor (LCA) as approximated using established taxonomic identity thresholds (Yarza *et al.* 2014, Vu *et al.* 2019) for each sequence is highlighted to aid in the identification of potentially novel taxa. The final taxonomic results table is exported as an interactive HTML object to allow for visual inspection prior to strain library creation.

2.4 Generating strain libraries

The third stage of analysis is performed *via* the "*isoLIB*" function, which facilitates the de-duplication of clonal or closely related isolates based on either their closest matching database hit (method = "closest_species") or agglomerative hierarchical-based clustering (Zou *et al.* 2020) of pairwise sequence similarities (method = "dark_mode"). The latter method is particularly effective for cataloging groups of novel microbes, i.e. dark matter, which may have close relatives within a given ecosystem or culture collection but otherwise lack reliable database sequence representation needed for accurate classification (Dueholm *et al.* 2020, 16). The default threshold value (0.995) uses a greedy incremental algorithm to iteratively assign sequences to different groups of which centroid sequence representatives have less than 99.5% pairwise similarity to one another (Rognes *et al.* 2016). Based on using 16S rRNA as a marker gene, this default threshold is expected to allow capture of intraspecies strain diversity in most communities since 98.7% pairwise similarity is the approximate cutoff for species demarcation (Kim *et al.* 2014). Alternatively, researchers seeking to develop a strain library with broad taxonomic coverage at the genus-to-species rank without regard for strain-level diversity may consider a threshold cutoff between 94.5% and 98.7% sequence similarity (Yarza *et al.* 2014). Ultimately, the generated strain library containing sequence information, metadata, and group designations is exported in table format as an interactive HTML object allowing visual inspection. If desired, the dataset can be further manually edited in place and saved as a CSV file. Users can also adjust parameters to specify a previously generated strain library. In this case, the new and old strain libraries are integrated, with priority given to the original strain groupings and designation of centroid sequence representatives in the older library. This process can be repeated as many times as necessary to support long-term and dynamic isolation workflows.

2.5 *In silico* benchmark analysis

To practically evaluate the *isoLIB* function in generating isolate libraries from human gut samples, we performed a simulated analysis on 6729 high-quality (HQ) genomes from the Inner Mongolian Gut Genome (IMGG) catalogue derived from fecal samples of $n = 60$ healthy adults (Jin *et al.* 2023). Briefly, *in silico* PCR was performed to extract genes of interest from each of the genomes (available here: https://figshare.com/articles/dataset/High_quality_NHMAG/19661523) using USEARCH's "search_pcr" command (Edgar et al. 2010) with three standard bacterial-identification primer sets targeting the full-length 16S rRNA gene (Waechter *et al.* 2023), V3–V6 regions of the 16S rRNA gene (Gloor *et al.* 2010), and the universal target (UT) of the *cpn60* gene (Vancuren and Hill 2019).

A total of 6533 genomes harboring all targets were further analyzed. In cases of intragenomic heterogeneity where a genome harbored multiple marker genes with different sequences, the consensus was determined using the DECIPHER package (Wright 2016) ("ConsensusSequence" command, ignoreNonBases = TRUE) to derive the most likely single sequence variant expected from Sanger sequencing-based identification during isolation workflows. Subsequently, each of the genomes (representative of distinct isolates in this simulation) were iteratively grouped based on marker gene sequence similarities of between 85% and 100% using isoLIB ("group_cutoff" = 0.85–1). Pooled library generation included all 6533 genomes whereas the individual library generations included genomes from a given individual's fecal sample [according to IMGG sample attributes described previously (Jin *et al.* 2023)]. For each type of library generation and marker gene evaluated, the percent of genome-level diversity captured at each cutoff was calculated as:

$$\text{Genome capture rate} = \frac{\text{Number of sequence groups generated by isoLIB}}{\text{Number of unique input genomes}} \qquad (1)$$

whereas the percent of species diversity captured was calculated as:

$$\text{Species capture rate} = \frac{\begin{array}{c}\text{Number of unique species represented}\\ \text{by isoLIB sequence groups}\end{array}}{\begin{array}{c}\text{Number of unique species}\\ \text{represented by input genomes}\end{array}} \qquad (2)$$

Accordingly, these ratios provide insight into how marker gene selection impacts the absolute number of strains that can be differentiated relative to the taxonomic coverage obtainable at each sequence similarity cutoff. As a point of reference for future culture collection initiatives, we also assessed the entire IMGG dataset [12 391 genomes, including those of varying completeness and quality (Jin *et al.* 2023)] as well as a larger dataset from the non-industrialized diverse microbiomes of Hadza hunter-gather populations [48 475 genomes from ultra-deep sequencing (Carter *et al.* 2023)] to comparatively evaluate how intraspecies genome diversity differs at the population and individual levels. The bash scripts used in the meta-analysis are available at: https://bdaisley.github.io/isolateR/benchmark/figure5_sh_script.html. The R scripts for visualizing meta-analysis results are available at: https://bdaisley.github.io/isolateR/benchmark/figure5_r_script.html.

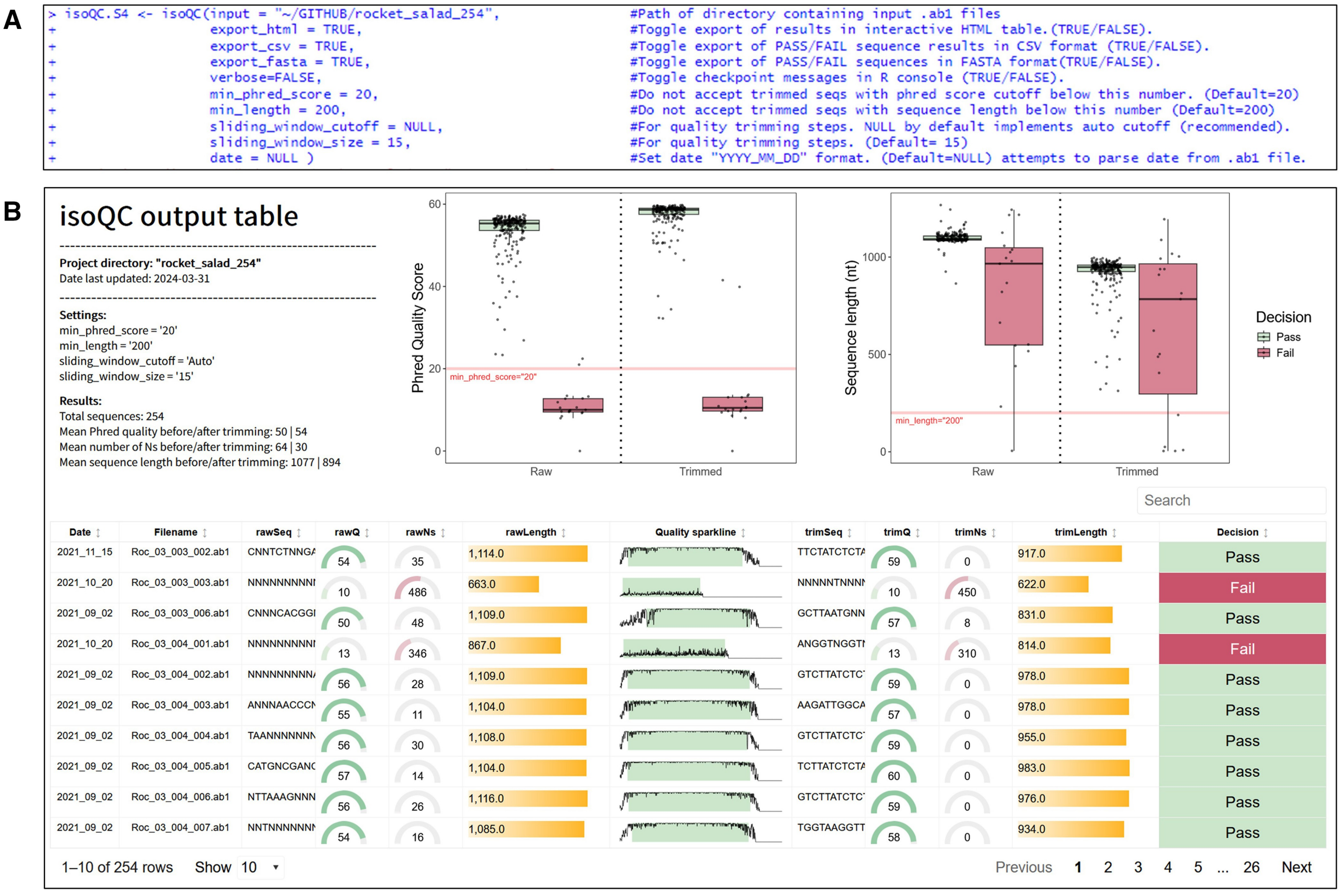
3 Results

To illustrate the usage of isolateR (Fig. 2), we analyze data from 254 bacterial isolates (16S rRNA gene Sanger sequence files downloaded from: https://dataverse.unimi.it/dataverse/vegmicroecol), which were originally derived from either conventional or organic ready-to-eat rocket salads found in local supermarkets across Milan, Italy (Mantegazza *et al.* 2023). This is a subset of the example data included in the isolateR package and the entire analysis including data import, quality trimming of sequences, taxonomic classification, and creation of the bacterial strain library can be completed by running three lines of R code. We discuss the execution and results of this workflow in more detail below.

3.1 Step 1: quality processing of Sanger sequences

Upon specifying the location of the directory containing the Sanger sequence files, the *isoQC* function can be run with default settings to automatically perform basecalling and adaptive quality trimming for all sequences in a single run (Fig. 2A). In the interactive HTML output (Fig. 2B), users can inspect descriptive statistics of the input sequences before and after trimming. The “Quality Sparkline” column further highlights the exact trimming region for each sequence, and the “Decision” column indicates whether a sequence should be discarded or not based on specified quality thresholds.

In this example, we see that the adaptive trimming approach in isolateR leads to a substantial improvement in the overall quality of sequence data. The mean length of input sequences is approximately 1077 bp before trimming and 894 bp after trimming. In tandem, there is a substantial reduction in the number of ambiguous bases decreasing from 64 to 30 with Phred quality scores increasing from an initial



```
> isoQC.S4 <- isoQC(input = "~/GITHUB/rocket_salad_254",          #Path of directory containing input .ab1 files
+                   export_html = TRUE,                           #Toggle export of results in interactive HTML table.(TRUE/FALSE).
+                   export_csv = TRUE,                            #Toggle export of PASS/FAIL sequence results in CSV format (TRUE/FALSE).
+                   export_fasta = TRUE,                          #Toggle export of PASS/FAIL sequences in FASTA format(TRUE/FALSE).
+                   verbose=FALSE,                                #Toggle checkpoint messages in R console (TRUE/FALSE).
+                   min_phred_score = 20,                         #Do not accept trimmed seqs with phred score cutoff below this number. (Default=20)
+                   min_length = 200,                             #Do not accept trimmed seqs with sequence length below this number (Default=200)
+                   sliding_window_cutoff = NULL,                 #For quality trimming steps. NULL by default implements auto cutoff (recommended).
+                   sliding_window_size = 15,                     #For quality trimming steps. (Default= 15)
+                   date = NULL )                                 #Set date "YYYY_MM_DD" format. (Default=NULL) attempts to parse date from .ab1 file.
```

isoQC output table

Project directory: "rocket_salad_254"
Date last updated: 2024-03-31

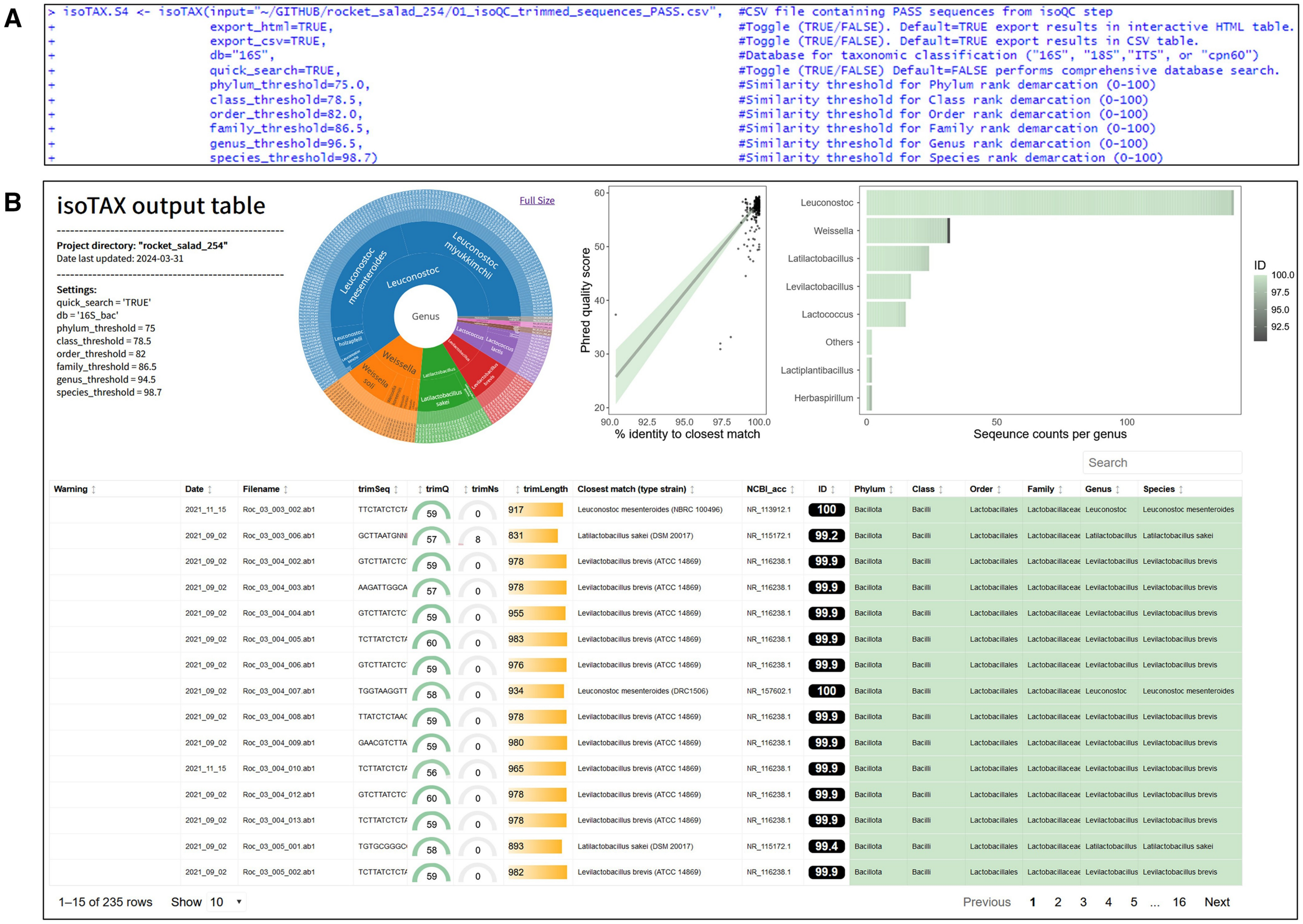
Settings:
min_phred_score = '20'
min_length = '200'
sliding_window_cutoff = 'Auto'
sliding_window_size = '15'

Results:
Total sequences: 254
Mean Phred quality before/after trimming: 50 | 54
Mean number of Ns before/after trimming: 64 | 30
Mean sequence length before/after trimming: 1077 | 894

Date	Filename	rawSeq	rawQ	rawNs	rawLength	Quality sparkline	trimSeq	trimQ	trimNs	trimLength	Decision
2021_11_15	Roc_03_003_002.ab1	CNNTCTNNGA	54	35	1,114.0		TTCTATCTCTA	59	0	917.0	Pass
2021_10_20	Roc_03_003_003.ab1	NNNNNNNNNN	10	486	663.0		NNNNNTNNNN	10	450	622.0	Fail
2021_09_02	Roc_03_003_006.ab1	CNNNCACGGI	50	48	1,109.0		GCTTAATGNN	57	8	831.0	Pass
2021_10_20	Roc_03_004_001.ab1	NNNNNNNNNN	13	346	867.0		ANGGTNGGTI	13	310	814.0	Fail
2021_09_02	Roc_03_004_002.ab1	NNNNNNNNNA	56	28	1,109.0		GTCTTATCTC	59	0	978.0	Pass
2021_09_02	Roc_03_004_003.ab1	ANNNAACCCN	55	11	1,104.0		AAGATTGGCA	57	0	978.0	Pass
2021_09_02	Roc_03_004_004.ab1	TAANNNNNNN	56	30	1,108.0		GTCTTATCTC	59	0	955.0	Pass
2021_09_02	Roc_03_004_005.ab1	CATGNCGANC	57	14	1,104.0		TCTTATCTCTA	60	0	983.0	Pass
2021_09_02	Roc_03_004_006.ab1	NTTAAAGNNN	56	26	1,116.0		GTCTTATCTC	59	0	976.0	Pass
2021_09_02	Roc_03_004_007.ab1	NNTNNNNNNN	54	16	1,085.0		TGGTAAGGTT	58	0	934.0	Pass

1–10 of 254 rows Show 10 ▾

Figure 2. isoQC function overview. (A) Input parameters. (B) Interactive HTML output table of results. The “Quality Sparkline” column shows the Phred quality over the entire sequence with the highlighted area representing the trimmed region of the sequence.



```
> isoTAX.S4 <- isoTAX(input="~/GITHUB/rocket_salad_254/01_isoQC_trimmed_sequences_PASS.csv",   #CSV file containing PASS sequences from isoQC step
+                     export_html=TRUE,                                                       #Toggle (TRUE/FALSE). Default=TRUE export results in interactive HTML table.
+                     export_csv=TRUE,                                                        #Toggle (TRUE/FALSE). Default=TRUE export results in CSV table.
+                     db="16S",                                                               #Database for taxonomic classification ("16S", "18S","ITS", or "cpn60")
+                     quick_search=TRUE,                                                      #Toggle (TRUE/FALSE) Default=FALSE performs comprehensive database search.
+                     phylum_threshold=75.0,                                                  #Similarity threshold for Phylum rank demarcation (0-100)
+                     class_threshold=78.5,                                                   #Similarity threshold for Class rank demarcation (0-100)
+                     order_threshold=82.0,                                                   #Similarity threshold for Order rank demarcation (0-100)
+                     family_threshold=86.5,                                                  #Similarity threshold for Family rank demarcation (0-100)
+                     genus_threshold=96.5,                                                   #Similarity threshold for Genus rank demarcation (0-100)
+                     species_threshold=98.7)                                                 #Similarity threshold for Species rank demarcation (0-100)
```

isoTAX output table

Project directory: "rocket_salad_254"
Date last updated: 2024-03-31

Settings:
quick_search = 'TRUE'
db = '16S_bac'
phylum_threshold = 75
class_threshold = 78.5
order_threshold = 82
family_threshold = 86.5
genus_threshold = 94.5
species_threshold = 98.7

Warning	Date	Filename	trimSeq	trimQ	trimNs	trimLength	Closest match (type strain)	NCBI_acc	ID	Phylum	Class	Order	Family	Genus	Species
	2021_11_15	Roc_03_003_002.ab1	TTCTATCTCTA	59	0	917	Leuconostoc mesenteroides (NBRC 100496)	NR_113912.1	100	Bacillota	Bacilli	Lactobacillales	Lactobacillaceae	Leuconostoc	Leuconostoc mesenteroides
	2021_09_02	Roc_03_003_006.ab1	GCTTAATGNNI	57	8	831	Latilactobacillus sakei (DSM 20017)	NR_115172.1	99.2	Bacillota	Bacilli	Lactobacillales	Lactobacillaceae	Latilactobacillus	Latilactobacillus sakei
	2021_09_02	Roc_03_004_002.ab1	GTCTTATCTC	59	0	978	Levilactobacillus brevis (ATCC 14869)	NR_116238.1	99.9	Bacillota	Bacilli	Lactobacillales	Lactobacillaceae	Levilactobacillus	Levilactobacillus brevis
	2021_09_02	Roc_03_004_003.ab1	AAGATTGGCA	57	0	978	Levilactobacillus brevis (ATCC 14869)	NR_116238.1	99.9	Bacillota	Bacilli	Lactobacillales	Lactobacillaceae	Levilactobacillus	Levilactobacillus brevis
	2021_09_02	Roc_03_004_004.ab1	GTCTTATCTC	59	0	955	Levilactobacillus brevis (ATCC 14869)	NR_116238.1	99.9	Bacillota	Bacilli	Lactobacillales	Lactobacillaceae	Levilactobacillus	Levilactobacillus brevis
	2021_09_02	Roc_03_004_005.ab1	TCTTATCTCTA	60	0	983	Levilactobacillus brevis (ATCC 14869)	NR_116238.1	99.9	Bacillota	Bacilli	Lactobacillales	Lactobacillaceae	Levilactobacillus	Levilactobacillus brevis
	2021_09_02	Roc_03_004_006.ab1	GTCTTATCTC	59	0	976	Levilactobacillus brevis (ATCC 14869)	NR_116238.1	99.9	Bacillota	Bacilli	Lactobacillales	Lactobacillaceae	Levilactobacillus	Levilactobacillus brevis
	2021_09_02	Roc_03_004_007.ab1	TGGTAAGGTT	58	0	934	Leuconostoc mesenteroides (DRC1506)	NR_157602.1	100	Bacillota	Bacilli	Lactobacillales	Lactobacillaceae	Leuconostoc	Leuconostoc mesenteroides
	2021_09_02	Roc_03_004_008.ab1	TTATCTCTAAC	59	0	978	Levilactobacillus brevis (ATCC 14869)	NR_116238.1	99.9	Bacillota	Bacilli	Lactobacillales	Lactobacillaceae	Levilactobacillus	Levilactobacillus brevis
	2021_09_02	Roc_03_004_009.ab1	GAACGTCTTA	59	0	980	Levilactobacillus brevis (ATCC 14869)	NR_116238.1	99.9	Bacillota	Bacilli	Lactobacillales	Lactobacillaceae	Levilactobacillus	Levilactobacillus brevis
	2021_11_15	Roc_03_004_010.ab1	TCTTATCTCTA	56	0	965	Levilactobacillus brevis (ATCC 14869)	NR_116238.1	99.9	Bacillota	Bacilli	Lactobacillales	Lactobacillaceae	Levilactobacillus	Levilactobacillus brevis
	2021_09_02	Roc_03_004_012.ab1	GTCTTATCTC	60	0	978	Levilactobacillus brevis (ATCC 14869)	NR_116238.1	99.9	Bacillota	Bacilli	Lactobacillales	Lactobacillaceae	Levilactobacillus	Levilactobacillus brevis
	2021_09_02	Roc_03_004_013.ab1	TCTTATCTCTA	59	0	978	Levilactobacillus brevis (ATCC 14869)	NR_116238.1	99.9	Bacillota	Bacilli	Lactobacillales	Lactobacillaceae	Levilactobacillus	Levilactobacillus brevis
	2021_09_02	Roc_03_005_001.ab1	TGTGCGGGC	58	0	893	Latilactobacillus sakei (DSM 20017)	NR_115172.1	99.4	Bacillota	Bacilli	Lactobacillales	Lactobacillaceae	Latilactobacillus	Latilactobacillus sakei
	2021_09_02	Roc_03_005_002.ab1	TCTTATCTCTA	59	0	982	Levilactobacillus brevis (ATCC 14869)	NR_116238.1	99.9	Bacillota	Bacilli	Lactobacillales	Lactobacillaceae	Levilactobacillus	Levilactobacillus brevis

1–15 of 235 rows Show 10 ▾

Previous 1 2 3 4 5 ... 16 Next

Figure 3. isoTAX function overview. (A) Input parameters. (B) Interactive HTML output table of results. "ID" column represents the pairwise global alignment sequence similarity of query sequences compared to the closest matching reference sequence from type strain database of interest. Taxonomic rank columns on right are highlighted based on the phylum- to species-level demarcation thresholds specified. Highlighted cells suggest confident taxonomic identity at each of the specified ranks, whereas unhighlighted cells are indicative of potentially novel taxa.

mean of approximately 50 to over 54 post-trimming (Fig. 2B).

3.2 Step 2: taxonomic classification of sequences

For common marker gene sequences such as 16S rRNA, taxonomic classification can be performed by running the *isoTAX* function with default settings. In the interactive HTML output (Fig. 3), the closest taxonomic matches and respective pairwise sequence identities are shown for each of the sequence representatives. Additionally, according to established taxonomic demarcation thresholds (Yarza *et al.* 2014), each of the taxonomic rank columns (phylum, class, order, family, genus, species) are shaded green based on whether sequence identities are supported at each level.

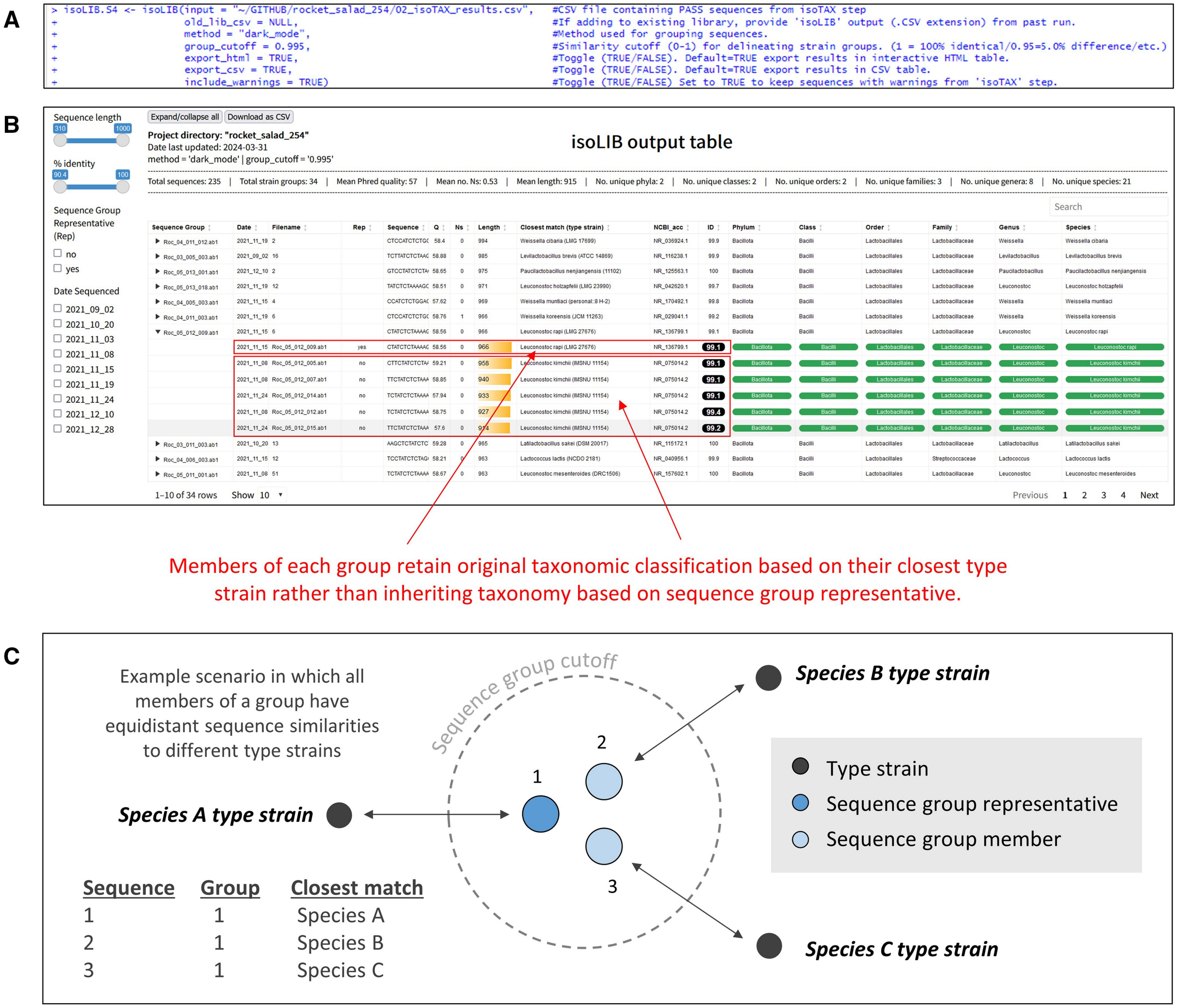
In the current example, we see several samples (Roc_04_005_005, Roc_04_006_001, Roc_04_006_005, Roc_04_006_012, and Roc_04_009_001) that have 100% sequence identities matching to *Lactococcus lactis* NCDO 604, and thus may represent clonal isolates. In contrast, there are two samples (Roc_03_013_001 and Roc_03_013_006) with sequence identities below the species demarcation threshold of 98.7%, suggesting these isolates may represent novel species found in association with rocket salad; these differences are visually reflected by an absence of green shading in the taxonomic species rank columns to help users rapidly identify potentially novel taxa of interest.

As an additional quality check, Phred quality scores for each sample are plotted against their respective sequence identities to the closest matching type strains, allowing outliers to be visually inspected (four samples in the current example with Phred scores less than 40) and removed from downstream analysis if applicable. Furthermore, a stacked bar plot and interactive sunburst chart are generated which highlight the most commonly detected taxa and their compositionality in the sample set, respectively (Fig. 3B).

3.3 Step 3: strain library creation

The final step involves generating a strain library using the *isoLIB* function (Fig. 4). This function enables the dereplication of clonal or closely related isolates based on customizable sequence similarity thresholds and offers flexible creation of libraries catering to specific research needs and the desired level of intraspecies or interspecies diversity. Interactive widgets in the HTML output document allow text-based searching as well as filtering of the strain library based on date, sequence length, pairwise identity, and other metadata.

In the example rocket salad culture collection (Fig. 4A and B), we see that the 235 sequences passing quality steps are



```
> isoLIB.S4 <- isoLIB(input = "~/GITHUB/rocket_salad_254/02_isoTAX_results.csv",   #CSV file containing PASS sequences from isoTAX step
+                     old_lib_csv = NULL,                                           #If adding to existing library, provide 'isoLIB' output (.CSV extension) from past run.
+                     method = "dark_mode",                                         #Method used for grouping sequences.
+                     group_cutoff = 0.995,                                         #Similarity cutoff (0-1) for delineating strain groups. (1 = 100% identical/0.95=5.0% difference/etc.)
+                     export_html = TRUE,                                           #Toggle (TRUE/FALSE). Default=TRUE export results in interactive HTML table.
+                     export_csv = TRUE,                                            #Toggle (TRUE/FALSE). Default=TRUE export results in CSV table.
+                     include_warnings = TRUE)                                      #Toggle (TRUE/FALSE) Set to TRUE to keep sequences with warnings from 'isoTAX' step.
```

Sequence Group Representative (Rep)

□ no
□ yes

Date Sequenced

□ 2021_09_02
□ 2021_10_20
□ 2021_11_03
□ 2021_11_08
□ 2021_11_15
□ 2021_11_19
□ 2021_11_24
□ 2021_12_10
□ 2021_12_28

Project directory: "rocket_salad_254"
Date last updated: 2024-03-31
method = 'dark_mode' | group_cutoff = '0.995'

isoLIB output table

Total sequences: 235 | Total strain groups: 34 | Mean Phred quality: 57 | Mean no. Ns: 0.53 | Mean length: 915 | No. unique phyla: 2 | No. unique classes: 2 | No. unique orders: 2 | No. unique families: 3 | No. unique genera: 8 | No. unique species: 21

Sequence Group	Date	Filename	Rep	Sequence	Q	Ns	Length	Closest match (type strain)	NCBI_acc	ID	Phylum	Class	Order	Family	Genus	Species
▶ Roc_04_011_012.ab1	2021_11_19	2		CTCCATCTCTGC	58.4	0	994	Weissella cibaria (LMG 17699)	NR_036924.1	99.9	Bacillota	Bacilli	Lactobacillales	Lactobacillaceae	Weissella	Weissella cibaria
▶ Roc_03_005_003.ab1	2021_09_02	16		TCTTATCTCTAAG	58.88	0	985	Levilactobacillus brevis (ATCC 14869)	NR_116238.1	99.9	Bacillota	Bacilli	Lactobacillales	Lactobacillaceae	Levilactobacillus	Levilactobacillus brevis
▶ Roc_05_013_001.ab1	2021_12_10	2		GTCCTATCTCTA	58.65	0	975	Paucilactobacillus nenjiangensis (11102)	NR_125563.1	100	Bacillota	Bacilli	Lactobacillales	Lactobacillaceae	Paucilactobacillus	Paucilactobacillus nenjiangensis
▶ Roc_05_013_018.ab1	2021_11_19	12		TATCTCTAAAAGC	58.51	0	971	Leuconostoc holzapfelii (LMG 23990)	NR_042620.1	99.7	Bacillota	Bacilli	Lactobacillales	Lactobacillaceae	Leuconostoc	Leuconostoc holzapfelii
▶ Roc_04_005_003.ab1	2021_11_15	4		CCATCTCTGGAC	57.62	0	969	Weissella muntiaci (personal::8 H-2)	NR_170492.1	99.8	Bacillota	Bacilli	Lactobacillales	Lactobacillaceae	Weissella	Weissella muntiaci
▶ Roc_04_011_003.ab1	2021_11_19	6		CTCCATCTCTGC	58.76	1	966	Weissella koreensis (JCM 11263)	NR_029041.1	99.2	Bacillota	Bacilli	Lactobacillales	Lactobacillaceae	Weissella	Weissella koreensis
▼ Roc_05_012_009.ab1	2021_11_15	6		CTATCTCTAAAAG	58.56	0	966	Leuconostoc rapi (LMG 27676)	NR_136799.1	99.1	Bacillota	Bacilli	Lactobacillales	Lactobacillaceae	Leuconostoc	Leuconostoc rapi
	2021_11_15	Roc_05_012_009.ab1	yes	CTATCTCTAAAAG	58.56	0	966	Leuconostoc rapi (LMG 27676)	NR_136799.1	99.1	Bacillota	Bacilli	Lactobacillales	Lactobacillaceae	Leuconostoc	Leuconostoc rapi
	2021_11_08	Roc_05_012_005.ab1	no	CTTCTATCTCTA	59.21	0	958	Leuconostoc kimchii (IMSNU 11154)	NR_075014.2	99.1	Bacillota	Bacilli	Lactobacillales	Lactobacillaceae	Leuconostoc	Leuconostoc kimchii
	2021_11_08	Roc_05_012_007.ab1	no	TTCTATCTCTAAA	58.85	0	940	Leuconostoc kimchii (IMSNU 11154)	NR_075014.2	99.1	Bacillota	Bacilli	Lactobacillales	Lactobacillaceae	Leuconostoc	Leuconostoc kimchii
	2021_11_24	Roc_05_012_014.ab1	no	TCTATCTCTAAAA	57.94	0	933	Leuconostoc kimchii (IMSNU 11154)	NR_075014.2	99.1	Bacillota	Bacilli	Lactobacillales	Lactobacillaceae	Leuconostoc	Leuconostoc kimchii
	2021_11_08	Roc_05_012_012.ab1	no	TCTATCTCTAAAA	58.75	0	927	Leuconostoc kimchii (IMSNU 11154)	NR_075014.2	99.4	Bacillota	Bacilli	Lactobacillales	Lactobacillaceae	Leuconostoc	Leuconostoc kimchii
	2021_11_24	Roc_05_012_015.ab1	no	TTCTATCTCTAAA	57.6	0	914	Leuconostoc kimchii (IMSNU 11154)	NR_075014.2	99.2	Bacillota	Bacilli	Lactobacillales	Lactobacillaceae	Leuconostoc	Leuconostoc kimchii
▶ Roc_03_011_003.ab1	2021_10_20	13		AAGCTCTATCTC	59.28	0	965	Latilactobacillus sakei (DSM 20017)	NR_115172.1	100	Bacillota	Bacilli	Lactobacillales	Lactobacillaceae	Latilactobacillus	Latilactobacillus sakei
▶ Roc_04_006_003.ab1	2021_11_15	12		TCCTATCTCTAG	58.21	0	963	Lactococcus lactis (NCDO 2181)	NR_040956.1	99.9	Bacillota	Bacilli	Lactobacillales	Streptococcaceae	Lactococcus	Lactococcus lactis
▶ Roc_05_011_001.ab1	2021_11_08	51		TCTATCTCTAAAA	58.67	0	963	Leuconostoc mesenteroides (DRC1506)	NR_157602.1	100	Bacillota	Bacilli	Lactobacillales	Lactobacillaceae	Leuconostoc	Leuconostoc mesenteroides

1–10 of 34 rows Show 10 ▾

Members of each group retain original taxonomic classification based on their closest type strain rather than inheriting taxonomy based on sequence group representative.

Example scenario in which all members of a group have equidistant sequence similarities to different type strains

Sequence	Group	Closest match
1	1	Species A
2	1	Species B
3	1	Species C

Figure 4. isoLIB function overview. (A) Input parameters. (B) Interactive HTML output table of results. Strain groups are based on sequence similarity at the specified threshold of interest. Members of each unique group are listed under the "filename" column and the sequence representatives (rep) are indicated in the "Rep" column. Taxonomic rank columns on right are highlighted based on the phylum- to species-level demarcation thresholds specified. Highlighted cells suggest confident taxonomic identity at each of the specified ranks, whereas unhighlighted cells are indicative of potentially novel taxa based on thresholds set. Interactive widgets allow text-based searching as well as filtering of the strain library based on date, sequence length, pairwise identity, and other metadata. (C) Schematic overview illustrating how sequences in the same group can be very closely related based on sequence identity yet have different taxonomic classifications based on closest matching type strain sequence identities.

represented by a total of 21 unique species and can be dereplicated to 34 strain groups based on the default grouping method ("dark_mode") and cutoff (0.995 = 99.5% sequence similarity). Each of the sequence group rows can be expanded *via* the dropdown menu to further inspect each of the samples within a given grouping. The top five groups by size are represented by *Leuconostoc miyukkimchii* (67), *Leuconostoc mesenteroides* (51), *Weissella soli* (17), *Levilactobacillus brevis* (16), and *Latilactobacillus sakei* (13).

Notably, de-replicating sequences at a given cutoff as in the "dark_mode" method often leads to the grouping of organisms with disparate taxonomies, a consequence of their equidistant sequence identities to closely matched type strains (Fig. 4C). This behavior aligns with an ecological species concept (Shapiro and Polz 2015), aiming to optimally organize distinct biological units according to their relevance in a specific habitat. For instance, in the rocket salad dataset, one of the samples, Roc_05_012_009, showing 99.1% identity to *Leuconostoc rapi*, groups with other samples (Roc_05_012_005, Roc_05_012_007, Roc_05_012_014, Roc_05_012_012, and Roc_05_012_015) identified closer to *Leuconostoc kimchii*, with identities ranging from 99.1% to 99.4% (Fig. 4C). Such grouping suggests these samples might represent a novel species or other biological unit meriting further exploration. Despite the apparent taxonomic diversity, the close relationship between these samples and their collective distinction from type strains nonetheless underscore the method's efficacy in de-replicating clonal or highly similar isolates. Transitioning from the ecological considerations enabled by "dark_mode", the *isoLIB* "closest_species" method addresses a different set of research objectives by enabling the creation of libraries where all members of a specific group are uniformly assigned the same taxonomic classification. This capability illustrates the platform's versatility in supporting a broad spectrum of study designs.

Emergent patterns of strain distribution and correlations with metadata attributes like sample source or collection date may further support hypothesis generation about ecological roles and interactions of specific taxa. Additionally, comparative analysis of strain libraries may reveal differences in strain diversity and abundance across habitats or conditions, providing valuable information for diagnostic or therapeutic applications.

3.4 Benchmark analysis and practical considerations

While the concept of a strain and the interpretation of species-level diversity of strains within microbiomes have been subject to debate (Van Rossum *et al.* 2020), adopting the perspective that any microbe with a distinct genome constitutes a strain allows for a broader and more inclusive understanding of microbial diversity. Here, we establish a real-case benchmark on how different *isoLIB* settings are expected to impact cataloging of strain-level diversity from the human gut by conducting an evaluation on 6533 HQ metagenomic assembled genomes (MAGs) obtained from 60 individuals from Inner Mongolia (Jin *et al.* 2023). Focusing on the typical scenario of using either full-length 16S rRNA, the V3–V6 regions 16S Rrna, or *cpn60* marker genes to identify bacterial isolates, we find that 3335, 2323, and 2224 MAGs, respectively, are distinguishable by at least a single nucleotide polymorphism or more compared to others within the pooled set of 6533 (Fig. 5A and B). This means that by applying a "group_cutoff = 1" for identical sequence dereplication in the *isoLIB* process, we could potentially capture about 34–51% of the overall strain diversity depending on the marker gene chosen. In contrast, using a "group_cutoff = 0.987"—reflecting the general species-level demarcation threshold for full-length 16S rRNA (Yarza *et al.* 2014)—results in capturing less than 15% of the strain diversity. However, this subset encompasses over 86% of the unique species found across the 60 individuals in the IMGG catalogue. At the same cutoff, the *cpn60* UT [previously shown to have greater sensitivity in demarcating phenotypically distinct sub-species and ecotypes (Vermette *et al.* 2010, Katyal *et al.* 2016)] captures a similar number of strains, which represent over 97% of species-level diversity. Together, these results align with other large-scale studies (Huttenhower *et al.* 2012, Gilbert *et al.* 2018, Pasolli *et al.* 2019), underscoring that strain-level variation within gut microbiomes is extensive at the population level and further highlighting the critical importance of choosing appropriate genetic markers and case-specific dereplication thresholds to accurately capture microbial diversity in isolation workflows.

Notably, strain-level resolution is greatly enhanced if we separately consider each of the gut microbiomes from the 60 Inner Mongolians as distinct habitats. For example, nearly double the number of MAGs (64–81% depending on marker gene) are distinguishable per individual at "group_cutoff = 1.0", and at the default setting ("group_cutoff = 0.995") all marker genes show high consistency in capturing approximately 61–64% of strain diversity and 97–99% of species-level diversity. Moreover, the rate at which decreasing the "group_cutoff" setting of *isoLIB* reduces the theoretical rate of strain capture is less rapid than when considering all strains pooled together across individuals (Fig. 5B). This phenomenon may be explained by the observation that conspecific strains sharing the same ecological niche frequently exhibit significant genomic divergence, which is attributable to variations in complementary accessory gene combinations underlying their co-existence (Hu *et al.* 2022). Exemplifying this point in relation to diseased host states, Wang *et al.* (2021) recently showed that over 70% of adherent-invasive *Escherichia coli* strains from inflammatory bowel disease patients could be efficiently tracked and quantified based on 16S rRNA identity—despite the fact that *E. coli* strains across individuals and environments are exceptionally challenging to differentiate based on any single gene marker (Hu *et al.* 2022). Thus, in cases of isolating microbes from habitats with high strain-level diversity (such as the human gut), it is advisable to generate libraries on a per individual host or habitat basis, to gain the greatest resolution in strain diversity.

As a point of reference further illustrating the effect of cataloging microbes from diverse habitats, we perform a secondary analysis on the non-industrialized microbiomes of Hadza hunter-gathers (Carter *et al.* 2023). Similar to the IMGG dataset, we show there are approximately 20 distinct MAGs detected per species unit across the 167 Hadza individuals, whereas each individual's microbiome is estimated to possess less than two distinct MAGs (strains) per species unit (Fig. 5C). Hadza individuals living in close proximity within specific "bush camps" are known to share microbes with one another extensively (Carter *et al.* 2023). Reflecting this unique lifestyle, we show that when individual microbiomes are pooled based on "bush camp", the number of MAGs per species unit increases only approximately ∼7, indicating a gradient zone where interindividual strain variability is less than expected. Nonetheless, both the IMGG and Hadza datasets support generating culture collections on a per-individual basis to capture the greatest strain-level diversity.

An important caveat is that these statistics are based on the assumption that all strains are culturable and have equal chances of being picked during isolation workflows. Since this is rarely the case, there are often tradeoffs when it comes to capturing the continuum of conspecific strain diversity relative to the overall taxonomic breadth attainable with consideration for throughput during isolation procedures. The default isoLIB parameter ("group_cutoff = 0.995") aims to strike a balanced compromise in terms of functionally grouping closely related strain variants, preventing clonal isolates, and ultimately expediting isolation workflow efficiency. Based on our analysis of the IMGG catalogue and the three primer sets tested, this default setting is estimated to capture 97–99% of bacterial species and approximately 61–64% of strain diversity when isolations are performed at the individual level. However, different taxonomic markers may be better or worse suited for delineating strains from a given habitat, and these estimates may not be generalizable to other domains of microbial life such as fungi and archaea, which have varying criteria for classifying species (Abarenkov *et al.* 2024). Thus, users may adjust this parameter higher or lower depending on desired study goals. These findings together serve as a useful benchmark for future culture collection endeavors and as a reminder of the inherent limitations of marker genes for delineating strain-level genetic diversity.

3.5 Comparisons to existing software

Table 1 details the similarities and differences between isolateR and other relevant software including

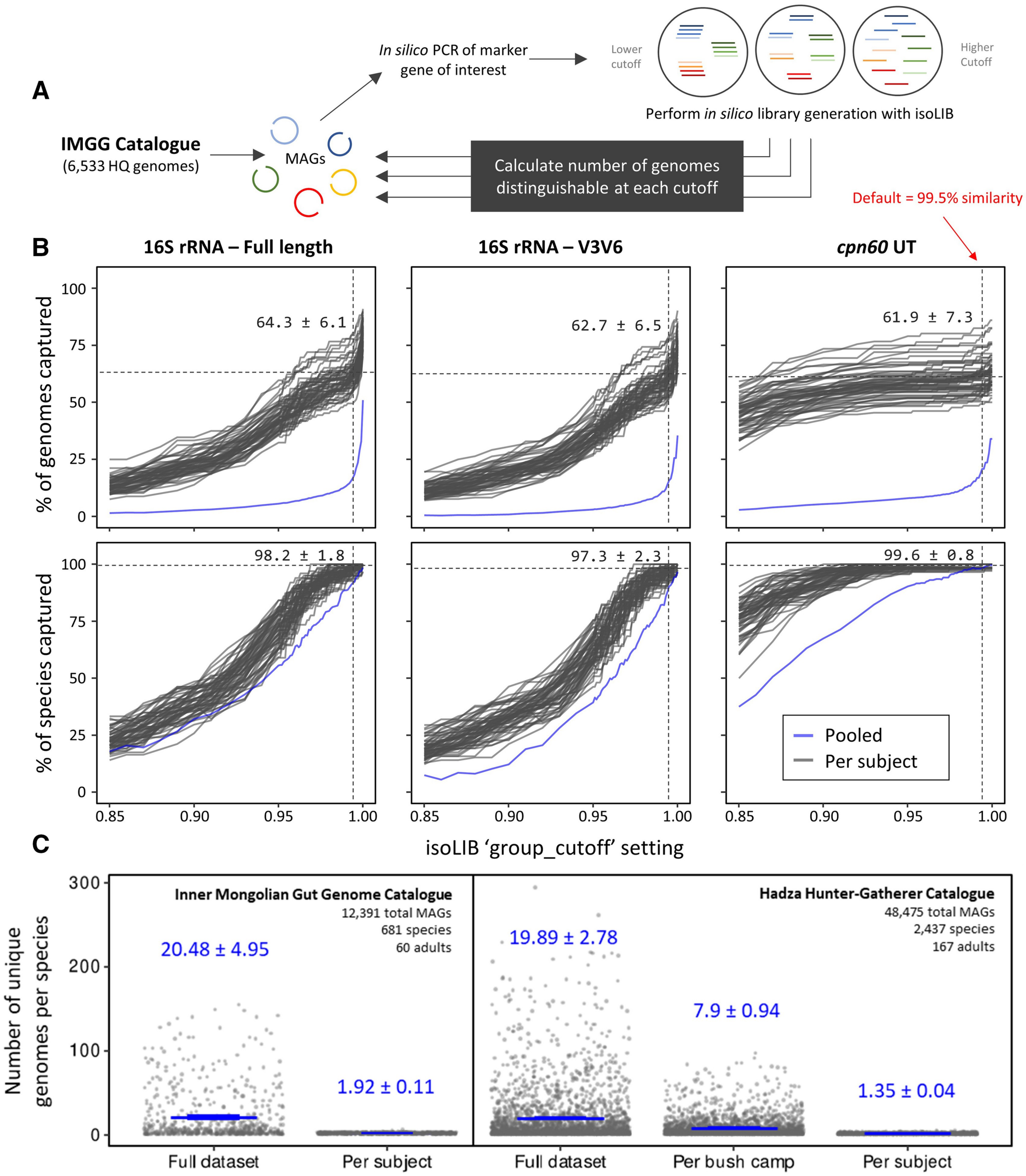


Figure 5. *In silico* benchmark of generating isolate libraries from the human gut. The *isoLIB* function aims to improve isolation workflow efficiency by grouping clonal or very closely related strains based on pairwise sequence identity of a given marker gene. Here, we assess a set of high-quality bacterial genomes from the Inner Mongolian Gut Genome (IMGG) catalogue to practically evaluate how the usage of different marker genes and *isoLIB* "group_cutoff" settings may impact diversity of culture collections from the human gut. (A) Schematic overview of methodology. A subset of 6533 high-quality (HQ) bacterial genomes from the IMGG catalogue, which contained both 16S rRNA and *cpn60* marker genes, were extracted *via in silico* PCR, and then iteratively grouped with *isoLIB* to simulate the dereplication of closely related strains based on sequence identity. (B) Results showing the percent of unique genomes relative to unique species expected to be captured at each cutoff when library generation is based on full-length 16S rRNA, V3–V6 regions 16S rRNA, or *cpn60* universal target (UT) marker genes. In every instance, capturing genome-level variation is enhanced by analyzing the metagenomes of individuals separately—underscoring the value of creating culture collections from individual niches as opposed to pooled samples. (C) Point of reference overview of the entire IMGG and Hadza Hunter–Gather catalogues showing intraspecies genome variability of the human gut at different scales.

sangeranalyserR (Chao *et al.* 2021), CAP3 (Huang and Madan 1999), DECIPHER (Wright 2016), Geneious (Kearse *et al.* 2012), MEGA7 (Kumar *et al.* 2016), SeqTrace (Stucky 2012), TraceTrack (Brazdilova *et al.* 2023), and LPSN API (Meier-Kolthoff *et al.* 2022).

Many software options provide trace visualization and sequence quality trimming, either manually or automatically (Stucky 2012, Chao *et al.* 2021, Brazdilova *et al.* 2023), but lack in taxonomic classification features, which are typically offered by a mutually exclusive set of specialized tools (Wright 2016). isolateR takes a significant step forward by automating taxonomic classification directly from Sanger sequencing input files, while also allowing dynamic reassignment of taxonomy to support the ever-evolving nature of hierarchical ranking systems in accordance with the rigorous standards of International Codes of Nomenclature (Meier-Kolthoff *et al.* 2022). Beyond processing and annotation features, isolateR is specifically tailored for generating

Table 1. Comparisons between isolateR and existing software.

Features	isolateR	Geneious	LPSN-API	sangeranalyseR	DECIPHER	CAP3	MEGA7	SeqTrace
Batch processing of Sanger files	X	X		X		X	X	X
Automated quality trimming of DNA sequences	X	X		X	X	X	X	X
Automated contig assembly	X	X		X	X	X	X	X
Phylogenetic tree output	X	X	X		X		X	
Taxonomic classification	X		X		X			
Dynamic re-assignment of taxonomy based on International Codes of Nomenclature	X		X					
Generation of sequence libraries	X							

sequence libraries that are instrumental in identifying novel taxa and in the efficient creation of large-scale microbial culture collections.

Ultimately, what sets isolateR apart is its ability to bridge the gap between functionalities of multiple bioinformatic tools, providing a comprehensive solution that addresses the varied needs of microbial genomics research. This integration simplifies microbial isolation workflows and unites the path from sequence acquisition to taxonomic classification and culture collection creation. Accordingly, these features position isolateR as a valuable tool for researchers focused on expanding our understanding of microbial diversity and streamlining the management of microbial resources.

4 Discussion

The entire isolateR workflow, from data import to the creation of the strain library can be completed efficiently using just three lines of R code. The processing time for the entire example dataset was under 15 min, demonstrating the software's capability to handle moderately large datasets with speed and accuracy within a standard laptop computing environment. In summary, isolateR is a robust and user-friendly tool supporting the high-throughput analysis of Sanger-based marker sequence data, comprehensive taxonomic classification, and straightforward creation of microbial strain libraries.

Data availability

The source code and data are available at https://github.com/bdaisley/isolateR.

Conflict of interest

E.A.V. is a co-founder of Nubiyota, a company focused on the human gut microbiome.

Funding

This work was supported by a Canada Research Chair Award (Canada Research Chairs Program, grant number: 232131 to E.A.V.), Canada First Research Excellence Fund (CFREF) Food from Thought (FfT) Thematic III Research Grant (THE3-00525 to E.A.V. and B.D.), and a Banting Postdoctoral Fellowship Award (BPF-489532-402947 to B.D.).

References

Abarenkov K, Nilsson RH, Larsson K-H *et al.* The UNITE database for molecular identification and taxonomic communication of fungi and other eukaryotes: sequences, taxa and classifications reconsidered. *Nucleic Acids Res* 2024;**52**:D791–7.

Altschul SF, Madden TL, Schäffer AA *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;**30**:2114–20.

Brazdilova K, Prihoda D, Ton Q *et al.* TraceTrack, an open-source software for batch processing, alignment and visualization of sanger sequencing chromatograms. *Bioinforma Adv* 2023;**3**:vbad083.

Carter MM, Olm MR, Merrill BD *et al.* Ultra-deep sequencing of Hadza hunter-gatherers recovers vanishing gut microbes. *Cell* 2023; **186**:3111–24.e13.

Case RJ, Boucher Y, Dahllöf I *et al.* Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Appl Environ Microbiol* 2007;**73**:278–88.

Chao K-H, Barton K, Palmer S *et al.* sangeranalyseR: simple and interactive processing of sanger sequencing data in R. *Genome Biol Evol* 2021;**13**:evab028.

Dueholm MS, Andersen KS, McIlroy SJ *et al.* Generation of comprehensive ecosystem-specific reference databases with species-level resolution by high-throughput full-length 16S rRNA gene sequencing and automated taxonomy assignment (AutoTax). *mBio* 2020; **11**:e01557-20.

Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 2010;**26**:2460–1.

Edgar RC. Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. *PeerJ* 2018;**6**:e4652.

Gilbert J, Blaser MJ, Caporaso JG *et al.* Current understanding of the human microbiome. *Nat Med* 2018;**24**:392–400.

Gloor GB, Hummelen R, Macklaim JM *et al.* Microbiome profiling by Illumina sequencing of combinatorial sequence-tagged PCR products. *PLoS One* 2010;**5**:e15406.

Hill JT, Demarest BL, Bisgrove BW *et al.* Poly peak parser: method and software for identification of unknown indels using sanger sequencing of polymerase chain reaction products. *Dev Dyn* 2014; **243**:1632–6.

Hu D, Fuller NR, Caterson ID *et al.* Single-gene long-read sequencing illuminates *Escherichia coli* strain dynamics in the human intestinal microbiome. *Cell Rep* 2022;**38**:110239.

Huang X, Madan A. CAP3: a DNA sequence assembly program. *Genome Res* 1999;**9**:868–77.

Huttenhower C, Gevers D, Knight R *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* 2012; **486**:207–14.

Janda JM, Abbott SL. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *J Clin Microbiol* 2007;**45**:2761–4.

Jin H, Quan K, He Q *et al.* A high-quality genome compendium of the human gut microbiome of Inner Mongolians. *Nat Microbiol* 2023; **8**:150–61.

Katyal I, Chaban B, Hill JE. Comparative genomics of cpn60-defined *Enterococcus hirae* ecotypes and relationship of gene content differences to competitive fitness. *Microb Ecol* 2016; **72**:917–30.

Kearse M, Moir R, Wilson A *et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 2012;**28**:1647–9.

Kim M, Oh H-S, Park S-C *et al.* Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int J Syst Evol Microbiol* 2014;**64**:346–51.

Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol* 2016; **33**:1870–4.

Lagier J-C, Khelaifia S, Alou MT *et al.* Culture of previously uncultured members of the human gut microbiota by culturomics. *Nat Microbiol* 2016;**1**:16203–8.

Mantegazza G, Gargari G, Duncan R *et al.* Ready-to-eat rocket salads as potential reservoir of bacteria for the human microbiome. *Microbiol Spectr* 2023;**11**:e02970-22.

Meier-Kolthoff JP, Carbasse JS, Peinado-Olarte RL *et al.* TYGS and LPSN: a database tandem for fast and reliable genome-based classification and nomenclature of prokaryotes. *Nucleic Acids Res* 2022; **50**:D801–7.

Nilsson RH, Larsson K-H, Taylor AFS *et al.* The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Res* 2019; **47**:D259–64.

Parte AC. LPSN—list of prokaryotic names with standing in nomenclature. *Nucleic Acids Res* 2014;**42**:D613–6.

Pasolli E, Asnicar F, Manara S *et al.* Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* 2019;**176**: 649–62.e20.

Rognes T, Flouri T, Nichols B *et al.* VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 2016;**4**:e2584.

Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 1977;**74**:5463–7.

Schoch CL, Seifert KA, Huhndorf S *et al.* Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for fungi. *Proc Natl Acad Sci* 2012;**109**:6241–6.

Shapiro BJ, Polz MF. Microbial speciation. *Cold Spring Harb Perspect Biol* 2015;**7**:a018143.

Stucky BJ. SeqTrace: a graphical tool for rapidly processing DNA sequencing chromatograms. *J Biomol Tech* 2012;**23**:90–3.

Tindall BJ, Rosselló-Móra R, Busse H-J *et al.* Notes on the characterization of prokaryote strains for taxonomic purposes. *Int J Syst Evol Microbiol* 2010;**60**:249–66.

Van Rossum T, Ferretti P, Maistrenko OM *et al.* Diversity within species: interpreting strains in microbiomes. *Nat Rev Microbiol* 2020; **18**:491–506.

Vancuren SJ, Hill JE. Update on cpnDB: a reference database of chaperonin sequences. *Database* 2019;**2019**:baz033.

Vermette CJ, Russell AH, Desai AR *et al.* Resolution of phenotypically distinct strains of *Enterococcus* spp. in a complex microbial community using cpn60 universal target sequencing. *Microb Ecol* 2010; **59**:14–24.

Vu D, Groenewald M, de Vries M *et al.* Large-scale generation and analysis of filamentous fungal DNA barcodes boosts coverage for kingdom fungi and reveals thresholds for fungal species and higher taxon delimitation. *Stud Mycol* 2019;**92**:135–54.

Waechter C, Fehse L, Welzel M *et al.* Comparative analysis of full-length 16S ribosomal RNA genome sequencing in human fecal samples using primer sets with different degrees of degeneracy. *Front Genet* 2023;**14**:1213829.

Wang J, Bleich RM, Zarmer S *et al.* Long-read sequencing to interrogate strain-level variation among adherent-invasive *Escherichia coli* isolated from human intestinal tissue. *PLoS One* 2021;**16**:e0259141.

Winter DJ. rentrez: an R package for the NCBI eUtils API. *R J* 2017; **9**:520–6.

Wright ES. Using DECIPHER v2.0 to analyze big biological sequence data in R. *R J* 2016;**8**:352–9.

Yarza P, Yilmaz P, Pruesse E *et al.* Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol* 2014;**12**:635–45.

Zou Q, Lin G, Jiang X *et al.* Sequence clustering in bioinformatics: an empirical study. *Brief Bioinform* 2020;**21**:1–10.

Bioinformatics, 2024, 40, 1–11
https://doi.org/10.1093/bioinformatics/btae448
Original Paper