Microbiome data science workflow

Leo Lahti, CSC course, Nov 28 – Dec 2, 2022

Overview of generic* metagenomics workflow

*This is generic; specific workflows can vary on the order of steps here and how they are done.

When working with your own data you should never follow any pipeline blindly. There can be critical differences based on your data.

sequencing facility → fastq files → demultiplex (split samples by barodes) → quality filter/trim (remove adapters/**primers**) → fasta files

might be done by sequencing facility

fastq files:
@HISEQ2500:282:1:1101:1220:1944 1
ATCGGATCG...
+
<G.<G<AGGII...

fastqc/multiqc

Some tools:
- sabre
- fastx_demux (usearch/vsearch)
- idemp
- fastx barcode splitter (fastx-toolkit)

Some tools:
- trimmomatic
- bbduk.sh (bbtools suite of tools)

fasta files:
>HISEQ2500:282:1:1101:1220:1944 1
ATCGGATCG...

no-assembly path

read-based analysis

Some tools:
- TIPP/SEPP
- metaphlan2
- humann2
- sourmash
- kraken

assembly path

consider testing assemblies with and w/o

digital normalization

Some tools:
- bbnorm
- diginorm

Count Table

	Sample_A	Sample_B	. . .
obj_1	0	428	. . .
obj_2	306	323	. . .
obj_3	217	1	. . .
. . .	. . .	. . .	. . .

Analysis

Some tools:
- phyloseq
- Breakaway
- DivNet
- CORNCOB
- SpiecEasi
- MaAsLin
- DESeq2

MetaQUAST is a great tool for comparing assemblies

(co)-assembly

map individual sample reads to (co)-assembly

Generate coverage information (mapping)

Some tools:
- bowtie2
- bwa

Some assemblers and tools:
- Megahit (assembler)
- SPAdes (assembler)
- idba-ud (assembler)
- MetAMOS (assembler and analysis pipeline)
- MetaCompass (reference-guided)
- MetagenomeScope (visualize assembly graphs)

Gene calling Functional/taxonomic profiling

Some tools:
- prodigal (identifies open reading frames)
- prokka (runs prodigal and performs annotations)
- GHOSTKOALA (web-hosted KEGG annotations)
- BLAST (protein nr db/refseq/COGs)

Recovering genomes from metagenomes

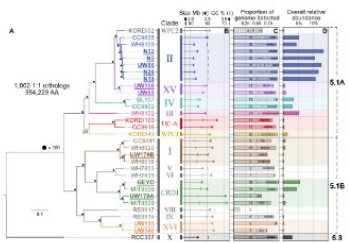
A note on MAGs:
MAGs (metagenome-assembled genomes) are not the same thing as isolate genomes. They are composite representative genomes of closely related genomic lineages.

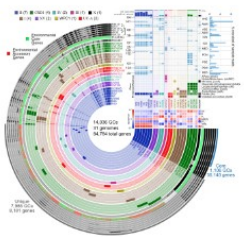
Some tools:
- anvi'o (interactive manual curation of bins; and much more)
- CONCOCT (kmer-based and coverage-based binning; also incorporated in anvi'o)
- COCACOLA (kmer-based, coverage-based, and incorporates paired-read linkage of contigs)
- MetaBAT2 (kmer-based and coverage-based binning tool)
- BinSanity (primarily coverage-based, optional second round kmer-based binning tool)
- checkm (genome-level taxonomy; and much more)
- DASTool (a tool for evaluating bins recovered by different methods)
- DESMAN (tool aimed at resolving strains)

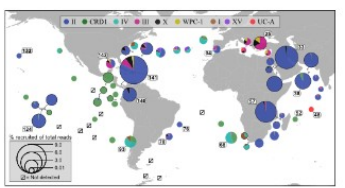
Some common genomics stuff

Phylogenomics
Comparative genomics
Pangenomics
Env. distributions

Some tools:
- anvi'o (integrated HMMs for common single-copy gene sets; integrated pangenomic workflow for identifying orthologs via OrthoMCL)
- PanOCT (identifies orthologs utilizing synteny information)
- StrainPhlAn/PanPhlAn (tools for strain-level analyses)
- MUSCLE (alignment software)
- FastTree (very fast, pseudo-maximum likelihood tree builder)
- RAxML (maximum likelihood tree builder)
- Mauve (whole-genome alignment)







astrobiomike.github.io

Happy Belly Bioinformatics

JOSE 10.21105/jose.00053

AstrobioMike

Orcid: 0000-0001-7750-9145

Lee, (2019). Happy Belly Bioinformatics: an open-source resource dedicated to helping biologists utilize bioinformatics. Journal of Open Source Education, 4(41), 53, https://doi.org/10.21105/jose.00053

Overview of generic* metagenomics workflow

*This is generic; specific workflows can vary on the order of steps here and how they are done.

When working with your own data you should never follow any pipeline blindly. There can be critical differences based on your data.

sequencing facility → fastq files → demultiplex (split samples by barodes) → quality filter/trim (remove adapters/primers) → fasta files

might be done by sequencing facility

@HISEQ2500:282:1:1101:1220:1944 1
ATCGGATCG...
+
<G.<G<AGGII...

Some tools:
- sabre
- fastx_demux (usearch/vsearch)
- idemp
- fastx barcode splitter (fastx-toolkit)

fastqc/multiqc

Some tools:
- trimmomatic
- bbduk.sh (bbtools suite of tools)

>HISEQ2500:282:1:1101:1220:1944 1
ATCGGATCG...

no-assembly path

read-based analysis

Some tools:
- TIPP/SEPP
- metaphlan2
- humann2
- sourmash
- kraken

Count Table

	Sample_A	Sample_B	. . .
obj_1	0	428	. . .
obj_2	306	323	. . .
obj_3	217	1	. . .
. . .	. . .	. . .	. . .

Analysis

Some tools:
- phyloseq
- Breakaway
- DivNet
- CORNCOB
- SpiecEasi
- MaAsLin
- DESeq2

assembly path

consider testing assemblies with and w/o

digital normalization

Some tools:
- bbnorm
- diginorm

MetaQUAST is a great tool for comparing assemblies

(co)-assembly

map individual sample reads to (co)-assembly

Generate coverage information (mapping)

Some tools:
- bowtie2
- bwa

Some assemblers and tools:
- Megahit (assembler)
- SPAdes (assembler)
- idba-ud (assembler)
- MetAMOS (assembler and analysis pipeline)
- MetaCompass (reference-guided)
- MetagenomeScope (visualize assembly graphs)

Gene calling Functional/taxonomic profiling

Recovering genomes from metagenomes → **A note on MAGs:** MAGs (metagenome-assembled genomes) are not the same thing as isolate genomes. They are composite representative genomes of closely related genomic lineages.

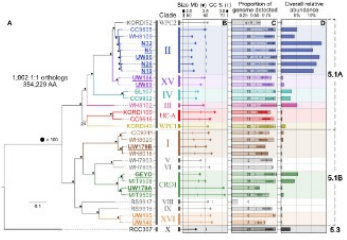
Some tools:
- prodigal (identifies open reading frames)
- prokka (runs prodigal and performs annotations)
- GHOSTKOALA (web-hosted KEGG annotations)
- BLAST (protein nr db/refseq/COGs)

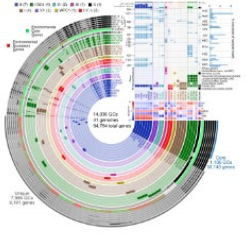
Some tools:
- anvi'o (interactive manual curation of bins; and much more)
- CONCOCT (kmer-based and coverage-based binning; also incorporated in anvi'o)
- COCACOLA (kmer-based, coverage-based, and incorporates paired-read linkage of contigs)
- MetaBAT2 (kmer-based and coverage-based binning tool)
- BinSanity (primarily coverage-based, optional second round kmer-based binning tool)
- checkm (genome-level taxonomy; and much more)
- DASTool (a tool for evaluating bins recovered by different methods)
- DESMAN (tool aimed at resolving strains)

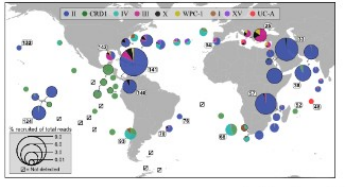
Some common genomics stuff

Phylogenomics
Comparative genomics
Pangenomics
Env. distributions

Some tools:
- anvi'o (integrated HMMs for common single-copy gene sets; integrated pangenomic workflow for identifying orthologs via OrthoMCL)
- PanOCT (identifies orthologs utilizing synteny information)
- StrainPhlAn/PanPhlAn (tools for strain-level analyses)
- MUSCLE (alignment software)
- FastTree (very fast, pseudo-maximum likelihood tree builder)
- RAxML (maximum likelihood tree builder)
- Mauve (whole-genome alignment)







astrobiomike.github.io

Our primary focus is on downstream analysis of taxonomic profiles of the microbial community, obtained for instance with:

- 16S rRNA amplicon sequencing
- (shallow) shotgun sequencing
- deep metagenome sequencing
- phylogenetic microarrays

Happy Belly Bioinformatics

JOSE 10.21105/jose.00053

AstrobioMike

Orcid: 0000-0001-7750-9145

Lee, (2019). Happy Belly Bioinformatics: an open-source resource dedicated to helping biologists utilize bioinformatics. Journal of Open Source Education, 4(41), 53, https://doi.org/10.21105/jose.00053

Common study designs

Case-control & Intervention

targeted experimental testing

Cross-sectional

population (cohort) studies

Prospective

long-term follow-ups

Longitudinal

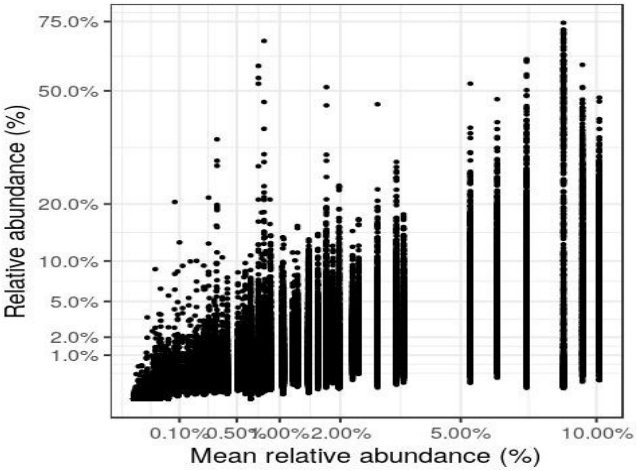
ecosystem dynamics

Microbiome data properties

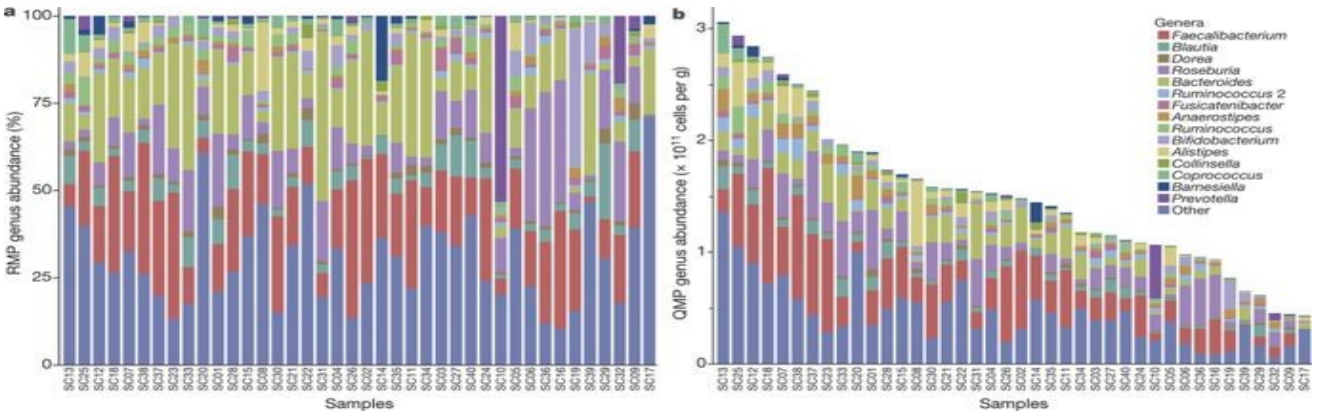
- Sparse
- Non-Gaussian
- Overdispersed
- Compositional
- Complex
- Stochastic
- Hierarchical

Heteroschedasticity

Data: HITChip Atlas (Lahti et al. Nature Communications 2014)

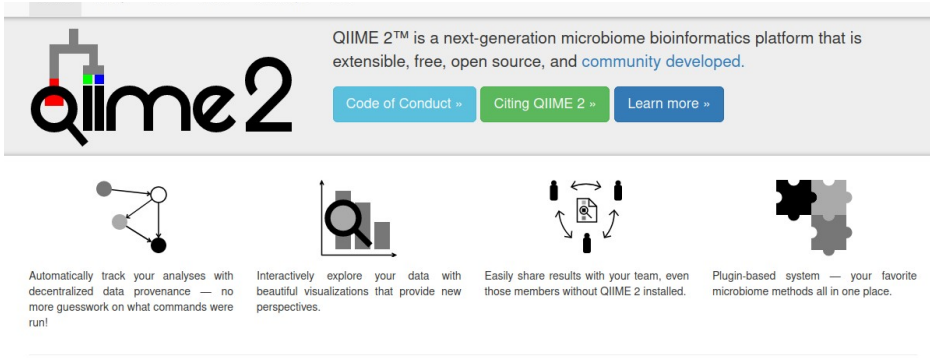


Compositionality



Vandeputte et al. Nature 551:507-511, 2017

open microbiome data science frameworks



Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

PeerJ ›

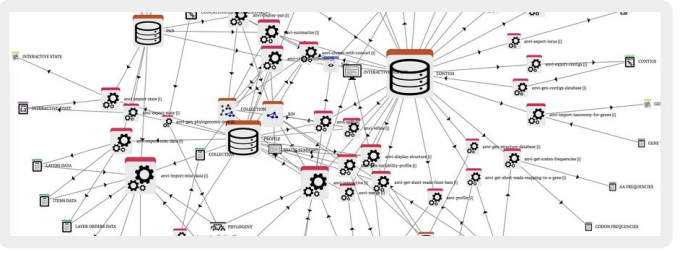
Anvi'o: an advanced analysis and visualization platform for 'omics data

Research article | Bioinformatics | Biotechnology | Computational Biology | Genomics | Microbiology

A. Murat Eren[1,2], Özcan C. Esen[1], Christopher Quince[3], Joseph H. Vineis[1], Hilary G. Morrison[1], Mitchell L. Sogin[1], Tom O. Delmont[1]

Published October 8, 2015

Anvi'o in a nutshell



Anvi'o is an open-source, community-driven analysis and visualization platform for 'omics data.

Workflow for community analysis

Data import

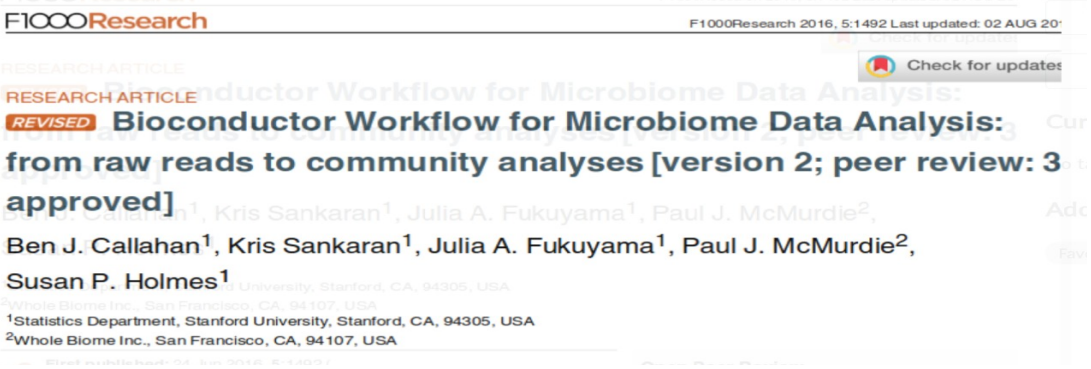
- Container preparation
- Sanity check (components)

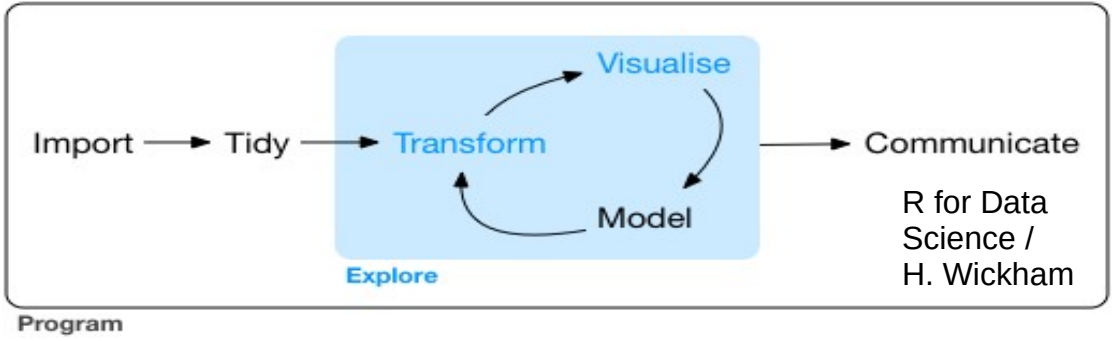
Data analysis

- Alpha & beta diversity
- Differential abundance
- Etc.

Reporting

- Summaries
- Reproducible reporting





Workflow for community analysis

Data import

- Container preparation
- Sanity check (components)

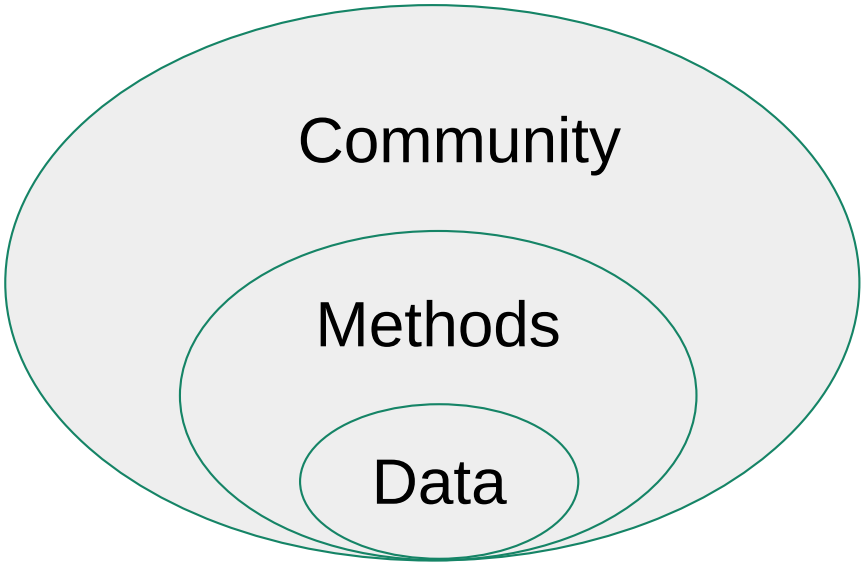
Data analysis

- Alpha & beta diversity
- Differential abundance
- Etc.

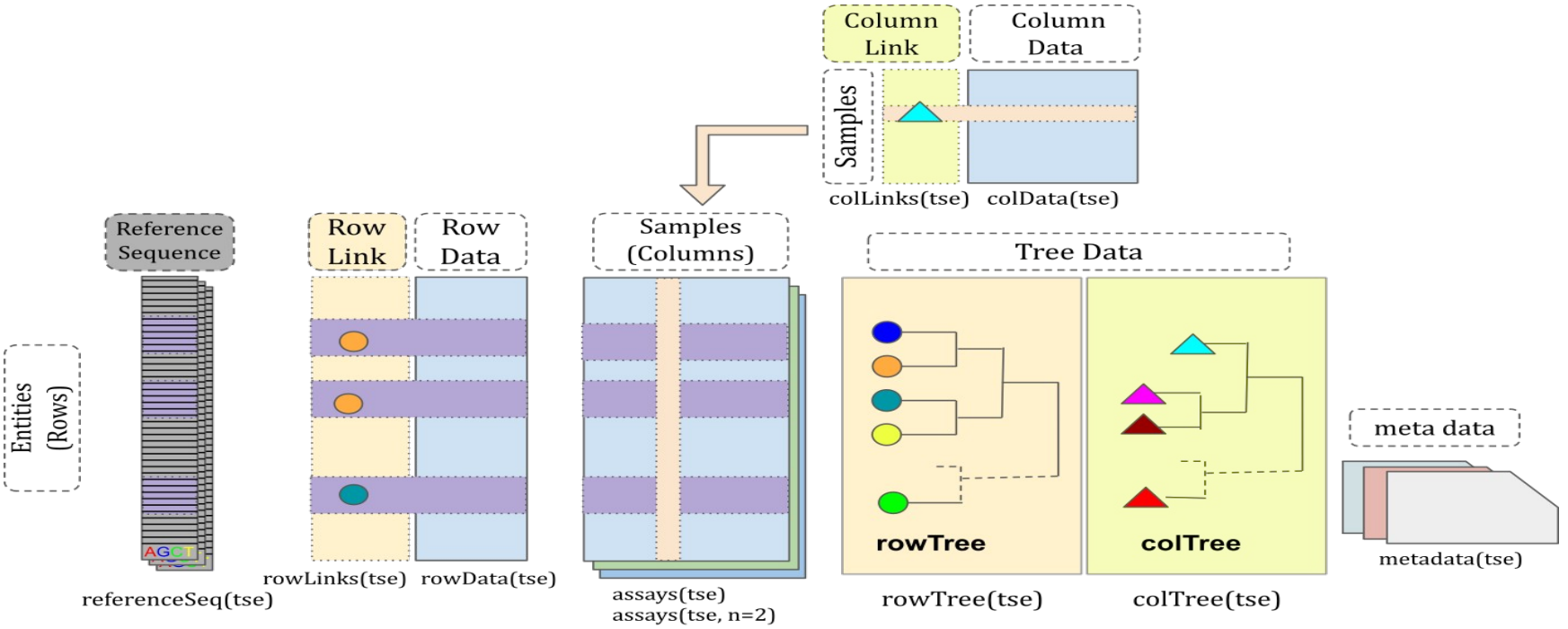
Reporting

- Summaries
- Reproducible reporting

Standardized data containers are central for the R/Bioc ecosystem



TreeSummarizedExperiment data container



Data analysis & visualization

Data import

- Container preparation
- Sanity check (components)

Data analysis

- Alpha & beta diversity
- Differential abundance
- Etc.

Reporting

- Summaries
- Reproducible reporting

```
plotColData(tse,
            "observed",
            "SampleType",
            colour_by = "Final_Barcode") +
    theme(axis.text.x = element_text(angle=45,hjust=1)) +
  ylab(expression(Richness[Observed]))
```

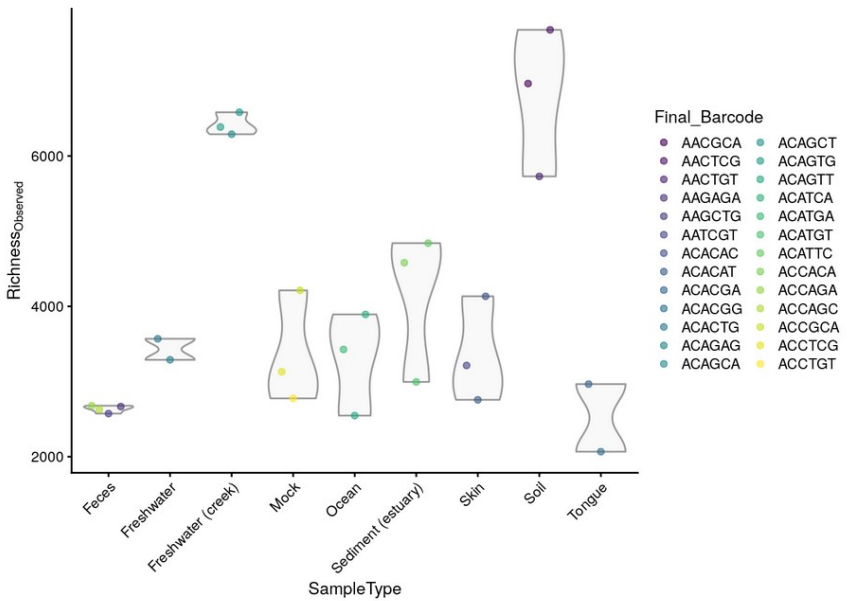


Figure 7.1: Shannon diversity estimates plotted grouped by sample type with colour-labeled barcode.

Typical tasks in community analysis

Standard:

- Community diversity (alpha diversity)
- Community similarity (beta diversity)
- Differential abundance analysis

Frequent:

- Community typing
- Co-occurrence networks
- Phylogenetic balances

```
plotRowTree(x[rowData(x)$Phylum %in% top_phyla_mean,],
            edge_colour_by = "Phylum",
            tip_colour_by = "prevalence",
            node_colour_by = "prevalence")
```

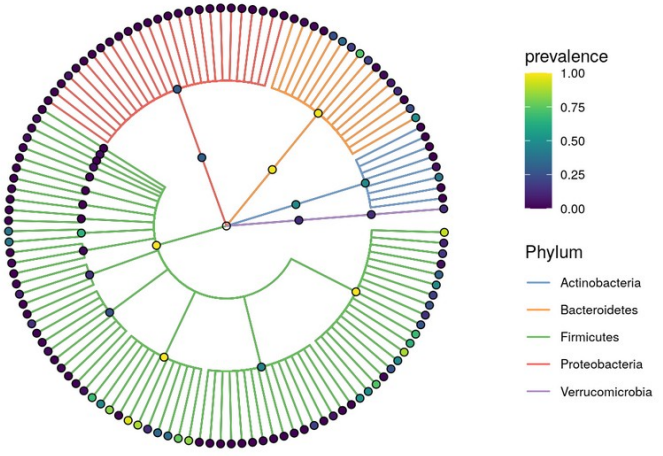
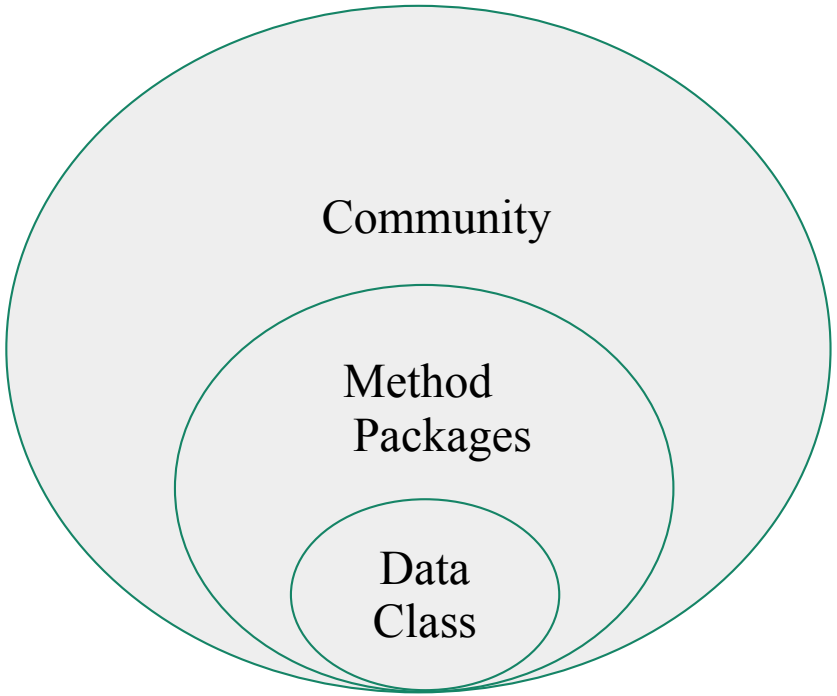


Figure 5.2: Prevalence of top phyla as judged by mean abundance

Reduce overlapping efforts, improve interoperability, ensure sustainability.



Data packages

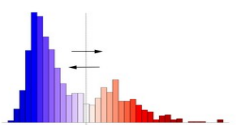
ExperimentHub

platforms all | rank 76 / 1974 | posts 2 / 1 / 2e+01 / 1 | in Bioc 4 years
build ok | updated before release | dependencies 72

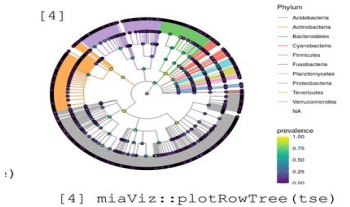
DOI: 10.18129/B9.bioc.ExperimentHub

mia – microbiome analysis

getDiversity(x)
calculateDMM(x)



miaViz - Visualization



Package ecosystem

Workflow for community analysis

Data import

- Container preparation
- Sanity check (components)

Data analysis

- Exploration
- Analysis & modeling

Reporting

- Summaries
- Reproducible reporting

Data containers support collaborative development of analysis methods & workflows

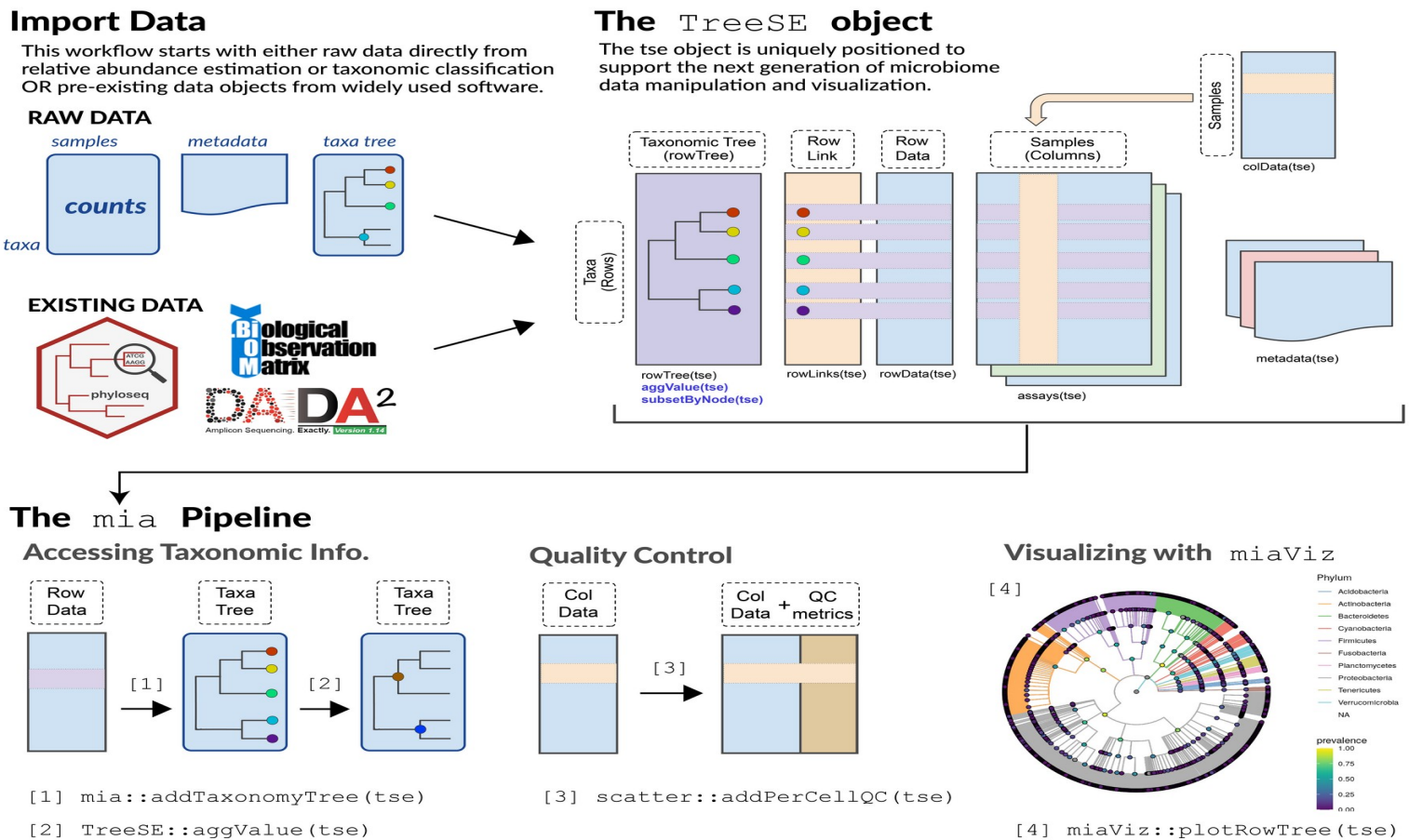


Fig: Domenick Braccia, EuroBioC 2020

microbiome.github.io

Orchestrating Microbiome Analysis with R and Bioconductor – online book: *beta version*

microbiome.github.io

Figure source: Moreno-Indias et al. (2021) Statistical and Machine Learning Techniques in Human Microbiome Studies: Contemporary Challenges and Solutions. Frontiers in Microbiology 12:11.

Microbiome Analysis

Authors: *Leo Lahti [aut], Sudarshan Shetty [aut], Felix GM Ernst [aut, cre]*
Version: *0.98.0003*
Modified: *2020-12-06*
Compiled: *2020-12-13*
Environment: *R version 4.0.0 (2020-04-24), Bioconductor 3.11*
License: *CC BY-NC-SA 3.0 US*
Copyright:
Source: *https://github.com/microbiome/MiaBook*

Preface

This website is a book on microbiome analysis in the Bioconductor universe and is showing common principles and workflows of performing microbiome analysis.

The book was borne out of necessity, while updating tools for microbiome analysis to work with common classes of the Bioconductor project handling count data of various sorts. It is heavily influenced by similar resources, such as the Orchestrating Single-Cell Analysis with Bioconductor book, phyloseq tutorials and microbiome tutorials.

We focus on microbiome analysis tools, new, updated and established methods. In the *Introduction* section, we show how to work with the key data infrastructure `TreeSummarizedExperiment` and related classes, how this framework relates to other infrastructure and how to load microbiome analysis data to work with in the context of this framework.

The second section, *Focus Topics*, is all about the steps for analyzing microbiome data, beginning with the most common steps and progressing to more specialized methods in subsequent sections.

The third section, *Appendix*, contains the rest of things we didn't find another place for, yet.

Acknowledgments

Course organizers:

- Finnish IT Center for Science (CSC)
- Department of Computing, University of Turku, Finland

Material preparation supported by: