

Data normalization and ecological analysis of microbiome data

BIOS² Workshop – Day 2

Steven Kembel – Sciences biologiques, UQAM



Outline

The nature of microbiome data

Data cleaning and exploration

Data normalization

Taxonomic composition of communities

Alpha diversity

Beta diversity

Differentially abundant taxa

The nature of microbiome data

After analyzing sequence data, we have the following matrices:

Error-corrected ASV abundances by sample - seqtab.nochim

	ASV_1	ASV_3	ASV_4	ASV_5	ASV_6	ASV_7	ASV_8	ASV_9	ASV_10	ASV_11
AUC.C1.2X1	1693	644	651	86	526	188	331	92	193	209
AUC.N1.3	973	145	776	0	61	9	14	196	172	36
AUC.N2.1X2	2138	696	26	798	239	162	3	402	81	559
AUC.N2.2	2721	510	97	5	689	119	15	33	266	55
AUC.S1.1.X2	3393	576	135	182	2	191	408	76	121	0
AUC.S1X1	2832	470	33	719	69	293	186	84	25	17
AUC.S2.2	977	422	823	29	364	584	161	347	127	69
BC.C1.2	3530	461	424	3	48	0	0	168	251	0
BC.C2.1X1	4070	123	314	212	44	19	0	12	145	8
BC.C2.3X2b	3852	139	19	6	31	42	1	19	8	1
BC.N1.3X2	2564	180	136	0	0	2	0	2	2	4
BC.N1	1046	182	87	132	15	0	14	33	159	40

Taxonomic annotations of each ASV - taxa.sp

	Kingdom	Phylum	Class	Order	Family	Genus
ASV_1	"Bacteria"	"Proteobacteria"	"Alphaproteobacteria"	"Rhizobiales"	"Beijerinckiaceae"	"Methylobacterium-Methylorubrum"
ASV_3	"Bacteria"	"Bacteroidota"	"Bacteroidia"	"Cytophagales"	"Hymenobacteraceae"	"Hymenobacter"
ASV_4	"Bacteria"	"Proteobacteria"	"Alphaproteobacteria"	"Sphingomonadales"	"Sphingomonadaceae"	"Sphingomonas"
ASV_5	"Bacteria"	"Proteobacteria"	"Alphaproteobacteria"	"Sphingomonadales"	"Sphingomonadaceae"	"Sphingomonas"
ASV_6	"Bacteria"	"Proteobacteria"	"Alphaproteobacteria"	"Rhizobiales"	"Beijerinckiaceae"	"Methylobacterium-Methylorubrum"
ASV_7	"Bacteria"	"Actinobacteriota"	"Actinobacteria"	"Micrococcales"	"Microbacteriaceae"	"Amnibacterium"
ASV_8	"Bacteria"	"Proteobacteria"	"Alphaproteobacteria"	"Sphingomonadales"	"Sphingomonadaceae"	"Sphingomonas"
ASV_9	"Bacteria"	"Actinobacteriota"	"Actinobacteria"	"Micrococcales"	"Microbacteriaceae"	"Frondihabitans"
ASV_10	"Bacteria"	"Proteobacteria"	"Alphaproteobacteria"	"Rhizobiales"	"Beijerinckiaceae"	"Methylobacterium-Methylorubrum"
ASV_11	"Bacteria"	"Proteobacteria"	"Alphaproteobacteria"	"Rhizobiales"	"Beijerinckiaceae"	"Methylobacterium-Methylorubrum"
ASV_12	"Bacteria"	"Proteobacteria"	"Alphaproteobacteria"	"Rhizobiales"	"Beijerinckiaceae"	"1174-901-12"
ASV_13	"Bacteria"	"Proteobacteria"	"Alphaproteobacteria"	"Rhizobiales"	"Beijerinckiaceae"	"Methylobacterium-Methylorubrum"
ASV_14	"Bacteria"	"Proteobacteria"	"Alphaproteobacteria"	"Sphingomonadales"	"Sphingomonadaceae"	"Sphingomonas"
ASV_15	"Bacteria"	"Proteobacteria"	"Alphaproteobacteria"	"Sphingomonadales"	"Sphingomonadaceae"	"Sphingomonas"
ASV_16	"Bacteria"	"Proteobacteria"	"Alphaproteobacteria"	"Sphingomonadales"	"Sphingomonadaceae"	"Sphingomonas"

Important caveats of microbiome data

Microbiome abundances are relative

- The total abundance of sequences in a sample is **unrelated** to the total abundance of organisms in communities

ASVs versus taxonomic annotations

- ASVs are based directly on the sequence data
- ASVs may be **missing taxonomic annotations** at different taxonomic ranks for several reasons
- **Be cautious** when analyzing the abundance of taxa at ranks other than ASVs (species, genus, family, etc.)

Important caveats of microbiome data

- Remember there are **numerous biases** and ASV abundance is an **imperfect measure of the abundance** of organisms in communities
- Complementary approaches and testing methods and assumptions can **help improve confidence** in estimates of microbiome structure

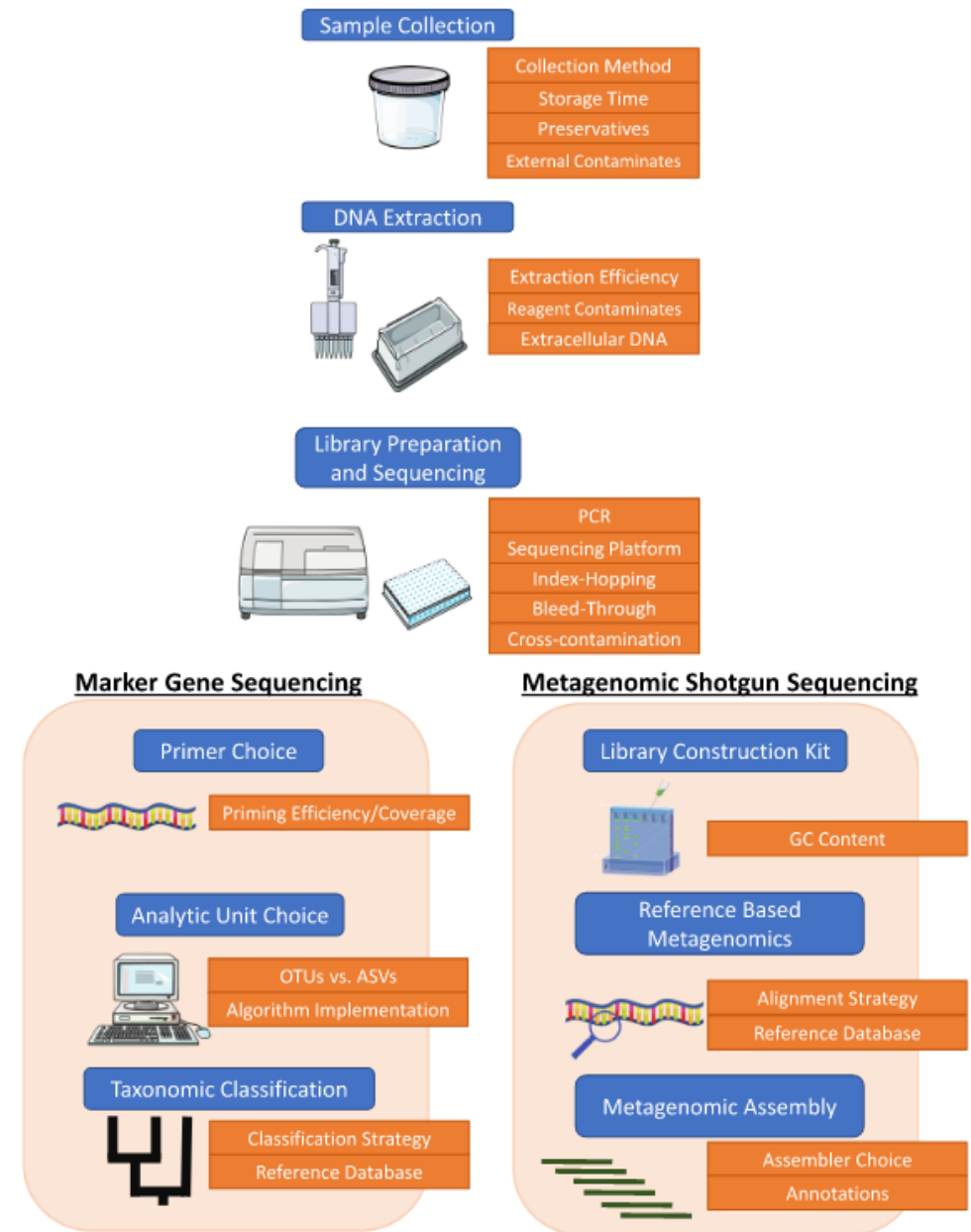


Fig. 1 The various stages that can introduce bias in sequenced-based human microbiome studies. Each blue box represents a stage in either DNA marker gene sequencing or DNA shotgun sequencing experiments. Orange boxes represent the various areas within a stage that can result in the introduction of systemic bias. Figure created using images from Servier Medical Art (<http://smart.servier.com>)

Ecological analysis of microbiome data

Data cleaning and exploration

Data normalization

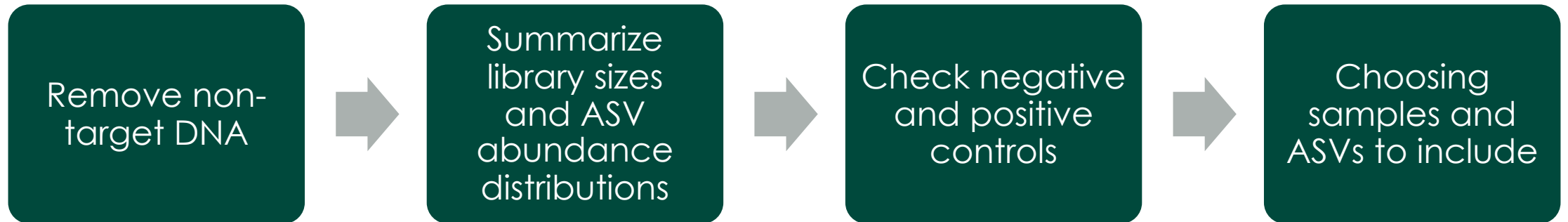
Taxonomic composition of communities

Alpha diversity

Beta diversity

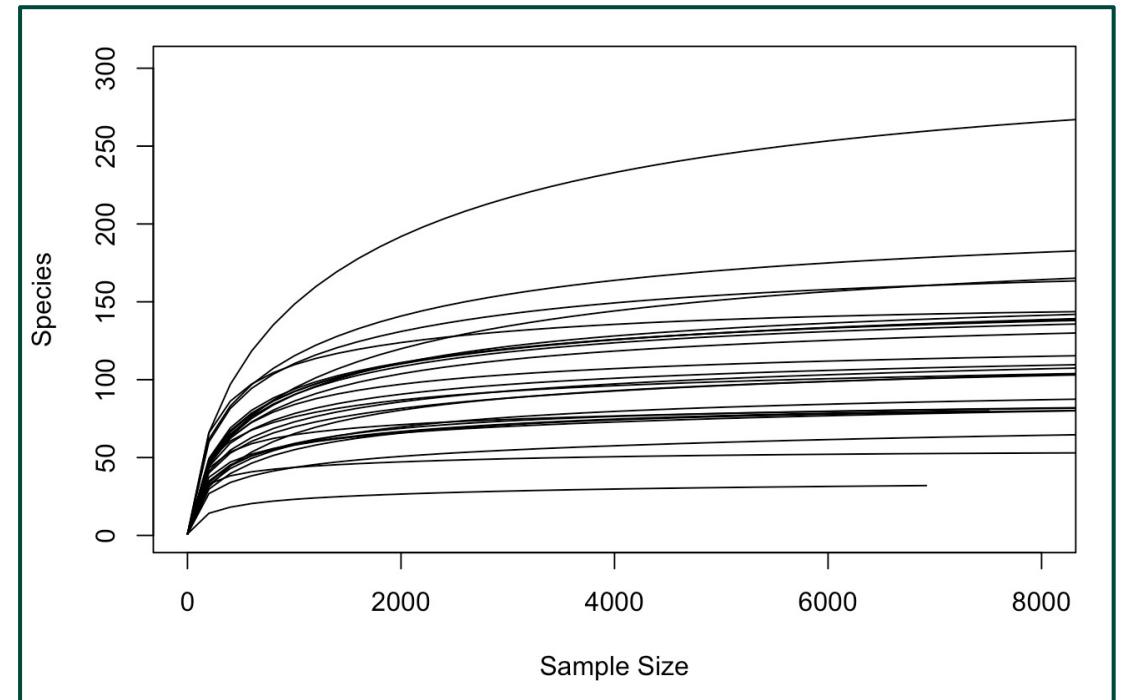
Differentially abundant taxa

Data cleaning and exploration



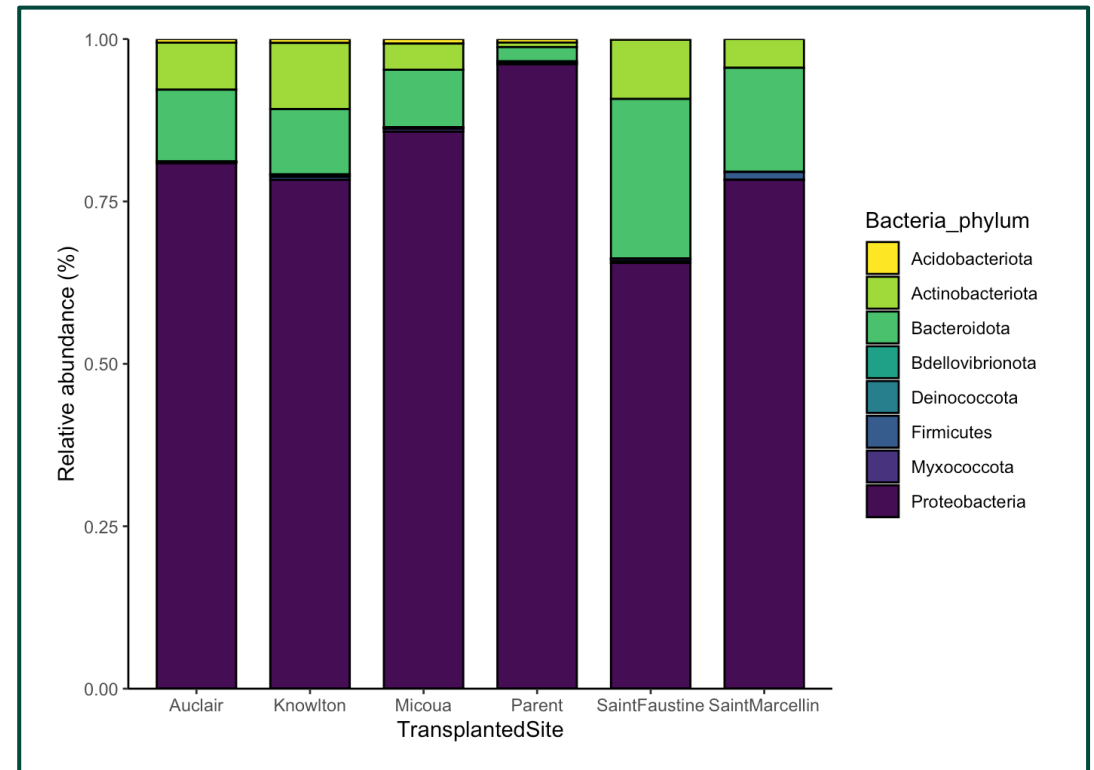
Data normalization

- Most measures of diversity are sensitive to sampling intensity
- We want to separate variation in diversity caused by some biological process from variation in diversity that is due to variation in library size
- Rarefaction is essential when calculating diversity metrics for microbiome data



Taxonomic composition of communities

- Limitations of taxonomic annotations
 - Reference databases are incomplete and biased
 - Ability to annotate is limited by sequence length
 - At finer taxonomic scales, many/most ASVs are missing annotations
 - Comparisons at those scales are missing many ASVs and sequences



Alpha diversity – within-sample diversity

Taxa richness

- Number of taxa in a sample

S

Shannon index

- Influenced by both the number of taxa in the sample (ASV richness) and the equitability of their abundances (evenness)

$$H = - \sum_{i=1}^S p_i \ln p_i$$

Measured with rarefied data

- Alpha diversity metrics are highly sensitive to sequencing depth and should always be computed on rarefied data

Beta diversity – between-sample diversity

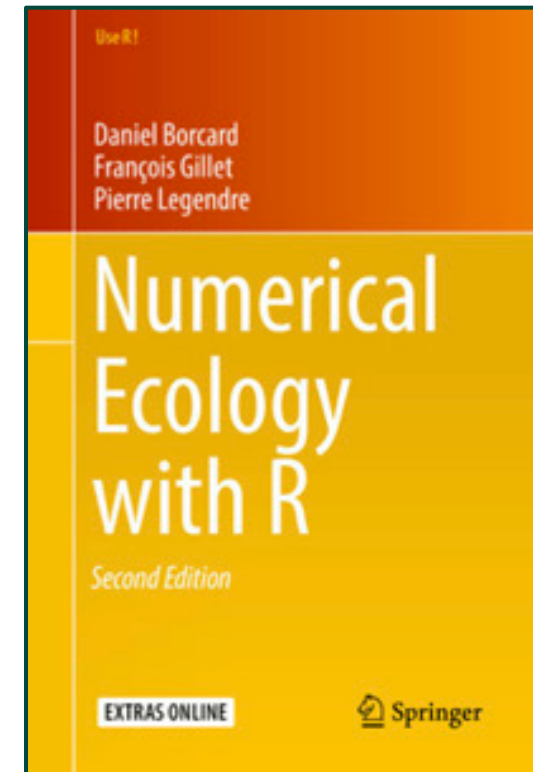
Beta diversity measures the **similarity of the composition** of two communities

- There are many beta-diversity metrics (a.k.a. community dissimilarity, distance metrics)
- Researchers have identified beta-diversity metrics and data transformations that perform well for the analysis of ecological communities (e.g. Legendre and Gallagher 2001 Oecologia):
 - Hellinger distances
 - Chord distances
 - Bray-Curtis distances

Beta diversity – between-sample diversity

Ordination methods

- They simplify complex multidimensional distances among communities into axes that capture major gradients of variation in community composition
- There are numerous ordination methods each with its own advantages and disadvantages
- Commonly used ordination approaches in microbial ecology
 - Principal Components Analysis (PCA)
 - Principal Coordinates Analysis (PCoA), a.k.a. Multidimensional Scaling (MDS)
 - Non-metric Multidimensional Scaling (NMDS)



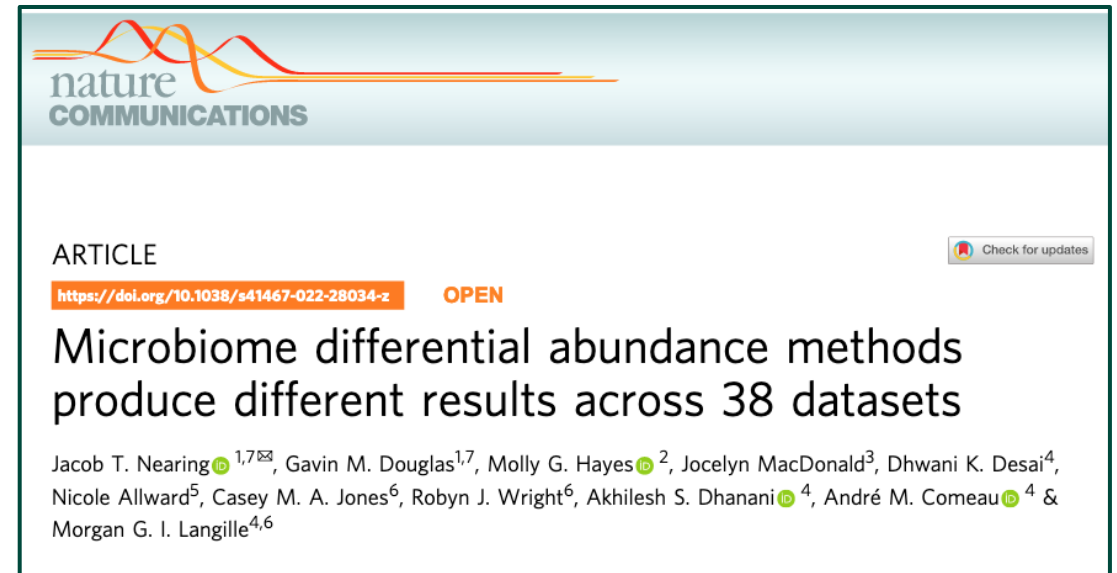
Beta diversity – between-sample diversity

PERMANOVA (Permutational multivariate analysis of variance)

- Non-parametric test of whether groups differ in their composition
- PERMANOVA can be used with any distance metric
- PERMANOVA can complement ordination analysis by testing the statistical significance of differences between groups that can be seen visually in ordination diagrams

Differentially abundant taxa

- There are numerous methods to test whether taxa are differentially abundant between groups of samples
- There is still debate about which methods work best (see Nearing et al. 2022. Nat. Comm)
- It's a good idea to try multiple approaches and compare results



The image shows a screenshot of a research article page from Nature Communications. At the top, the Nature Communications logo is displayed with a stylized orange and red wave graphic. Below the logo, the word "ARTICLE" is written in a light blue font. To the right of "ARTICLE" is a "Check for updates" button. Below this, the article's DOI is shown in a blue box: <https://doi.org/10.1038/s41467-022-28034-z>, followed by the word "OPEN" in orange. The main title of the article is "Microbiome differential abundance methods produce different results across 38 datasets" in a large, bold, black font. Below the title, the authors are listed: Jacob T. Nearing (ORCID iD), Gavin M. Douglas^{1,7}, Molly G. Hayes (ORCID iD)², Jocelyn MacDonald³, Dhvani K. Desai⁴, Nicole Allward⁵, Casey M. A. Jones⁶, Robyn J. Wright⁶, Akhilesh S. Dhanani (ORCID iD)⁴, André M. Comeau (ORCID iD)⁴ & Morgan G. I. Langille^{4,6}.