Differential Abundance Analysis (DAA)

July 28 2024

Phuc Loi Luu, PhD

Email: Luu.p.loi@googlemail.com

Content

- Review previous lectures
- Read paper “The Follicular Skin Microbiome in Patients With Hidradenitis Suppurativa and Healthy Controls”
- What is differential abundance (DA) analysis?
- How to do DAA?
- Tools
- Problems
- Projects for the end term
- Hands on
  - DSeq2: Khải
  - ANCOM-BC1/2: DS Trà

Review previous lectures

Following initial quality assessment/control steps, such as primer(s) removal, demultiplexing and quality filtering, the 16S amplicon sequences are either clustered into Operational Taxonomic Units (OTUs) representing the common working definition of bacterial species[9] by OTU picking algorithms (e.g. **<1>**), or grouped into **<2>** using denoising algorithms (e.g. **<3>** and **<4>**). After the construction of OTU or **<2>**, these observed counts are typically organized into a large matrix referred to as the **<5>** table. Some researchers or software packages such as **<6>** represent samples by columns and features (OTUs or **<2>**) by rows, but this representation is not universal.

Review previous lectures

- Following initial quality assessment/control steps, such as primer(s) removal, demultiplexing and quality filtering, the 16S amplicon sequences are either clustered into Operational Taxonomic Units (OTUs) representing the common working definition of bacterial species[9] by OTU picking algorithms (e.g. VSEARCH, UPARSE[10]), or grouped into Sequence Variants (SVs) using denoising algorithms (e.g. DADA2[11] and Deblur[12]). After the construction of OTU or SV, these observed counts are typically organized into a large matrix referred to as the feature table. Some researchers or software packages such as QIIME2[13] represent samples by columns and features (OTUs or SVs) by rows, but this representation is not universal.

https://www.nature.com/articles/s41522-020-00160-w#Sec1

Original Investigation

September 2017

The Follicular Skin Microbiome in Patients With Hidradenitis Suppurativa and Healthy Controls

Hans Christian Ring, MD[1]; Jonathan Thorsen, MD[2]; Ditte M. Saunte, MD, PhD[1]; et al

» Author Affiliations | Article Information

JAMA Dermatol. 2017;153(9):897-905. doi:10.1001/jamadermatol.2017.0904

Key Points

Question Does the cutaneous microbiome in hidradenitis suppurativa differ from that in healthy controls?

Findings In this case-control study that included 30 patients with hidradenitis suppurativa and 24 healthy controls, next-generation sequencing analysis demonstrated a significantly different microbiome in patients with hidradenitis suppurativa (lesional and nonlesional) compared with that in healthy controls.

Meaning Overall, the data suggest that a dysbiotic microbiome may have a role in the pathogenesis of hidradenitis suppurativa.

https://jamanetwork.com/journals/jamadermatology/fullarticle/2628754

Table. Background Characteristics of Study Participants

Characteristic	Patients With HS	Healthy Controls
Total analyzed	30	24
Female, No. (%)	19 (63)	13 (54)
Age, mean (SD), y[a]	46.9 (14.0)	32.2 (12.0)
White race/ethnicity, No. (%)	30 (100)	23 (96)
BMI, mean (SD)[a]	30.5 (7.7)	26.5 (2.6)
Smoking status, No. (%)[b]	20 (67)	7 (29)
Sartorius score, mean (SD)	24 (20)	NA
HS severity, No. (%)		
Hurley stage 1	0	NA
Hurley stage 2	26 (87)	NA
Hurley stage 3	4 (13)	NA
Duration of lesions, median (IQR), d	21 (14-60)	NA
Location of lesions	Groin (n = 15), axilla (n = 15)	Axilla
Diameter of lesions, mean (SD), cm	2.76 (1.38)	NA

Abbreviations: BMI, body mass index (calculated as weight in kilograms divided by height in meters squared); HS, hidradenitis suppurativa; IQR, interquartile range; NA, not applicable.

[a] $P < .05$ by Wilcoxon rank sum test.

[b] $P < .05$ by χ^2 test.

Figure 1. Distribution of the Top 10 Most Abundant Species Found in Patients With Hidradenitis Suppurativa (HS) and in Healthy Controls

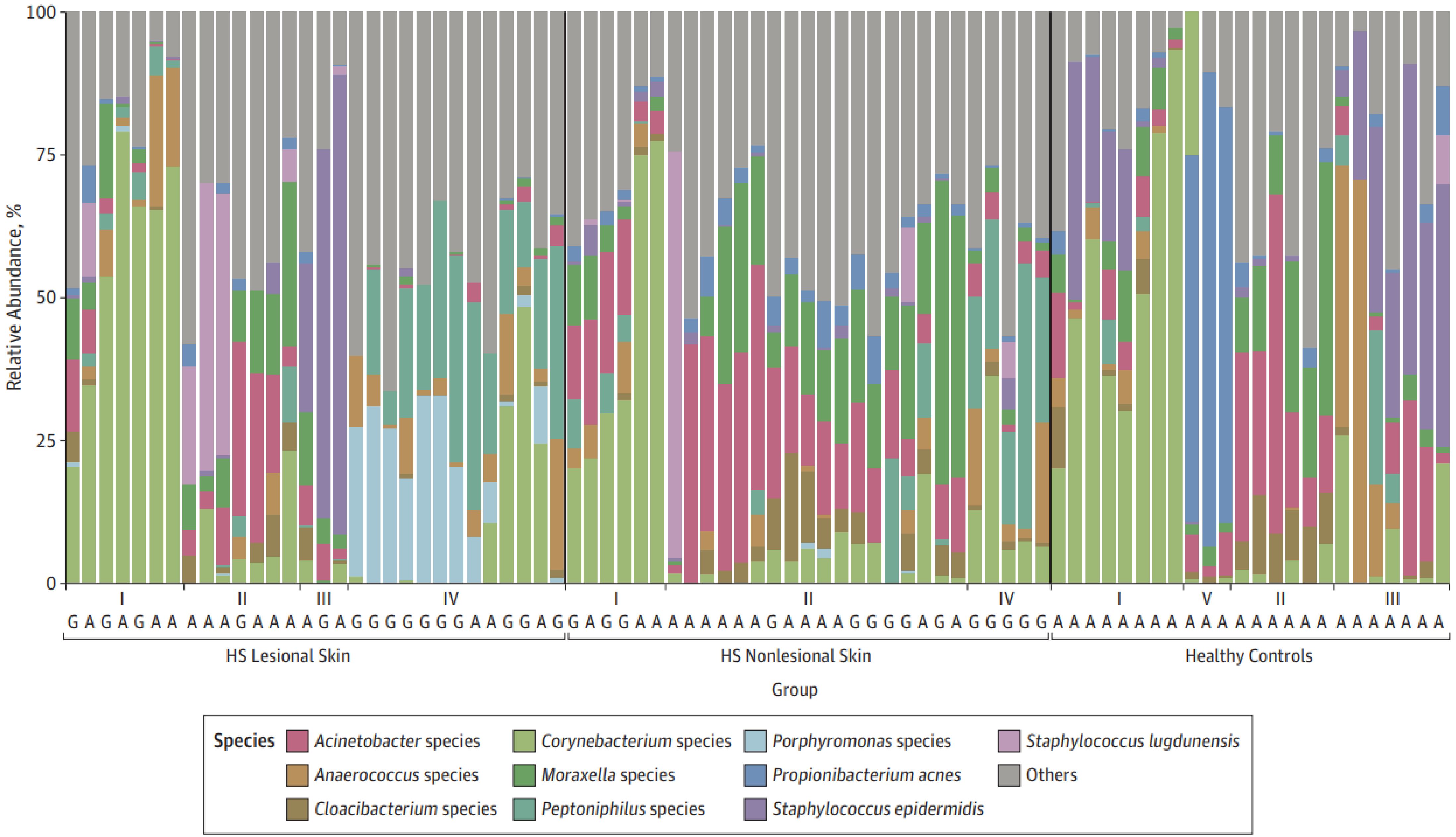
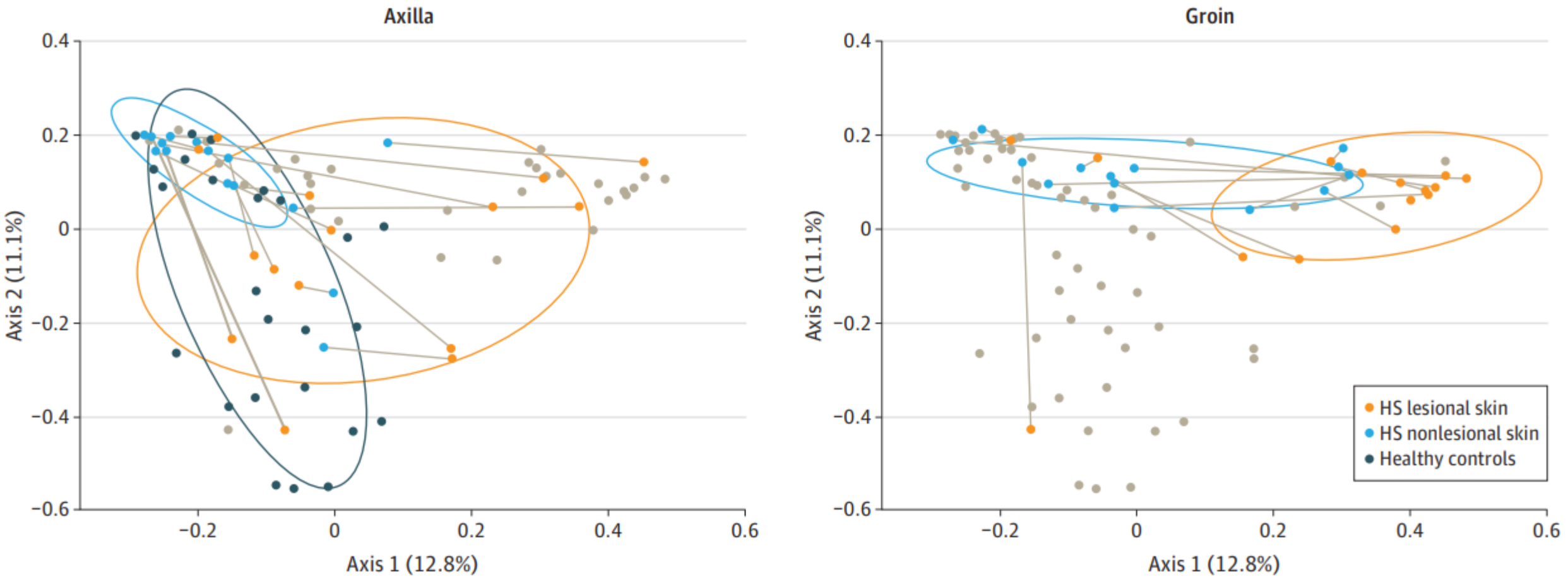
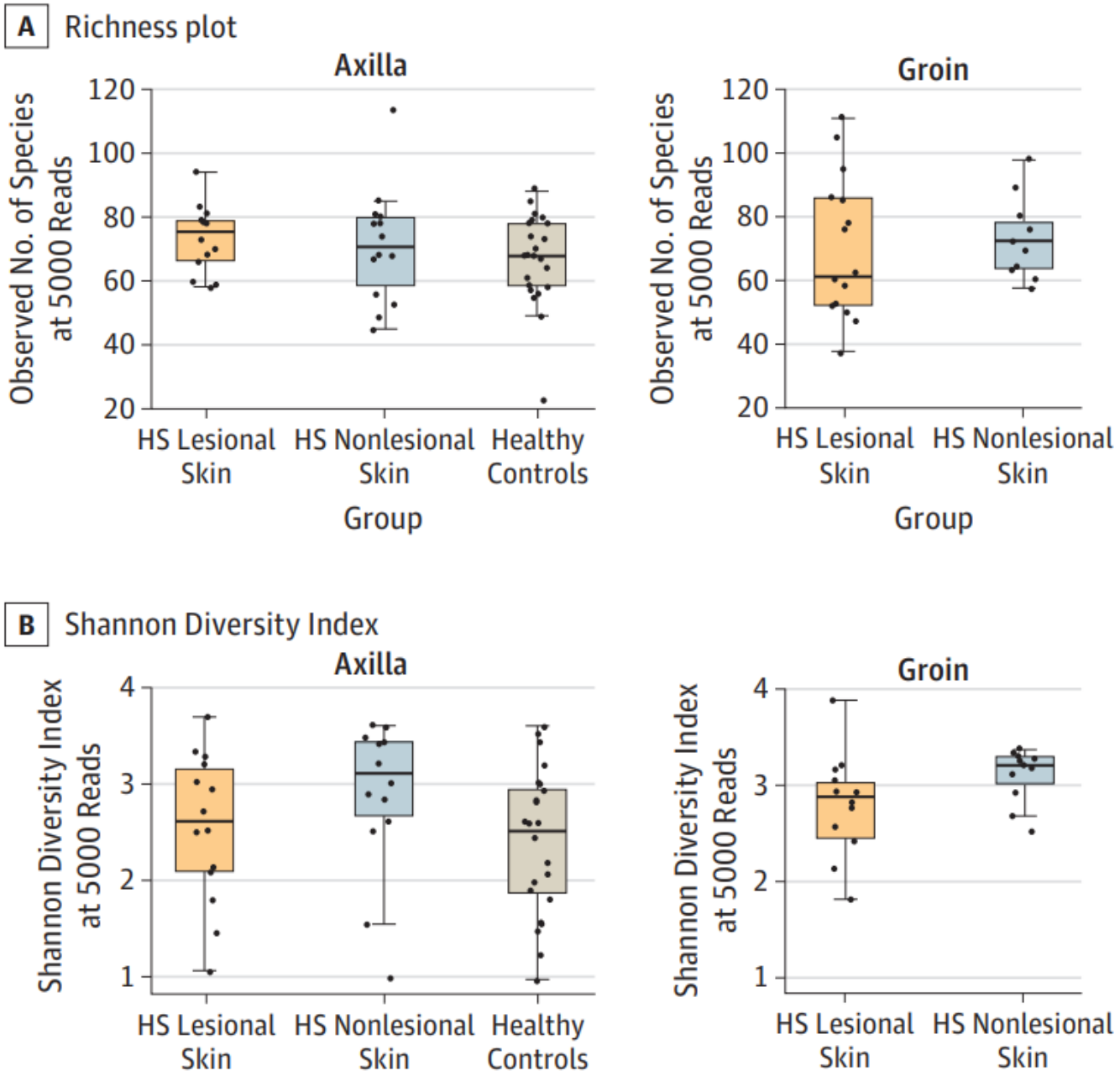


Figure 2. Bray-Curtis Principal Coordinates Analysis Plot Showing Differences Between the 3 Groups



Shown are differences between the 3 groups (HS lesional skin, HS nonlesional skin, and healthy controls) in axillary and groin samples. Each lesional sample is connected with its corresponding nonlesional sample. Gray dots indicate samples from the opposite anatomical location (axilla or groin). Ellipses indicate the 75% prediction areas of samples from each group. HS indicates hidradenitis suppurativa.

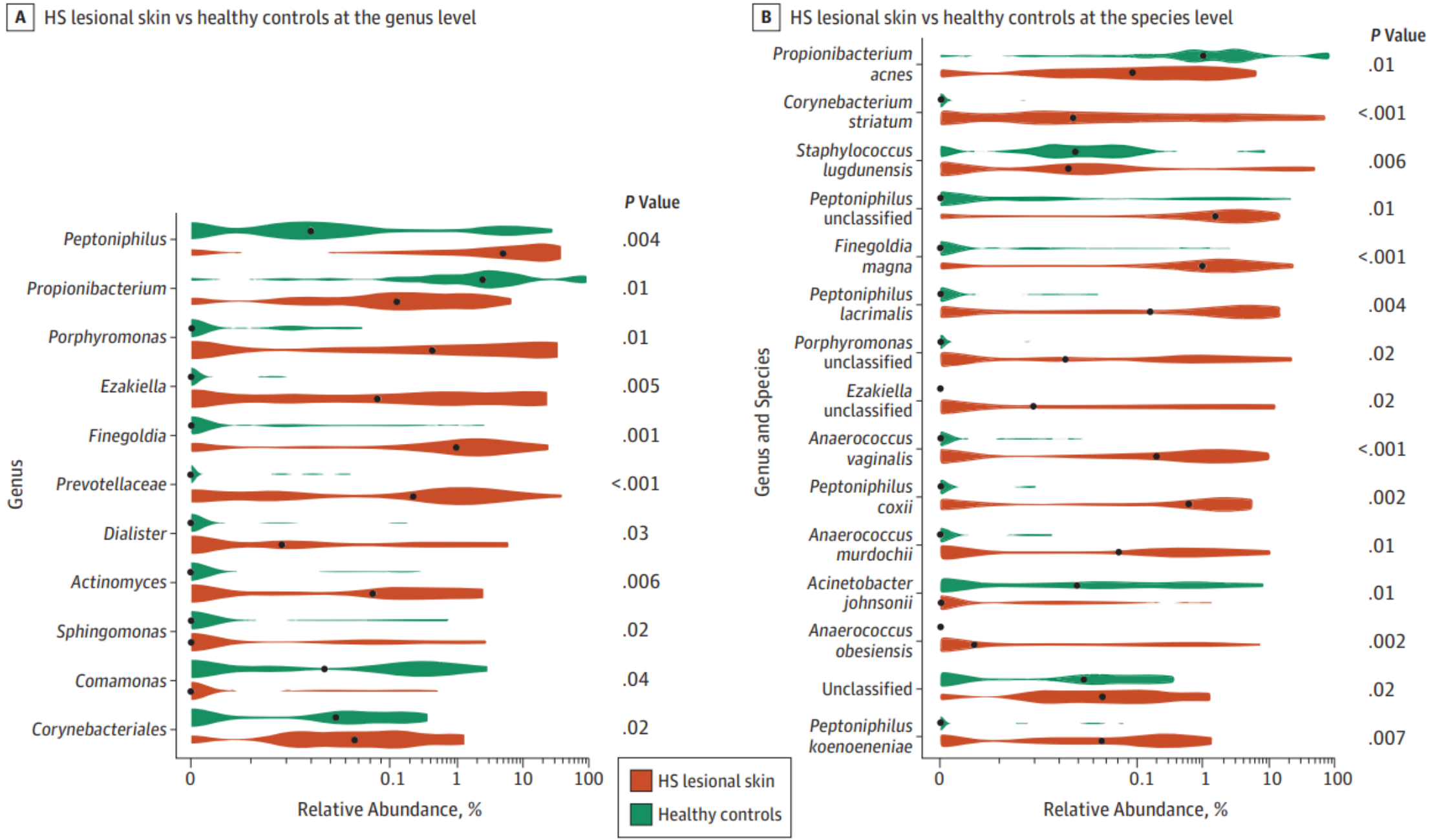
Figure 3. Richness Plot and Shannon Diversity Index



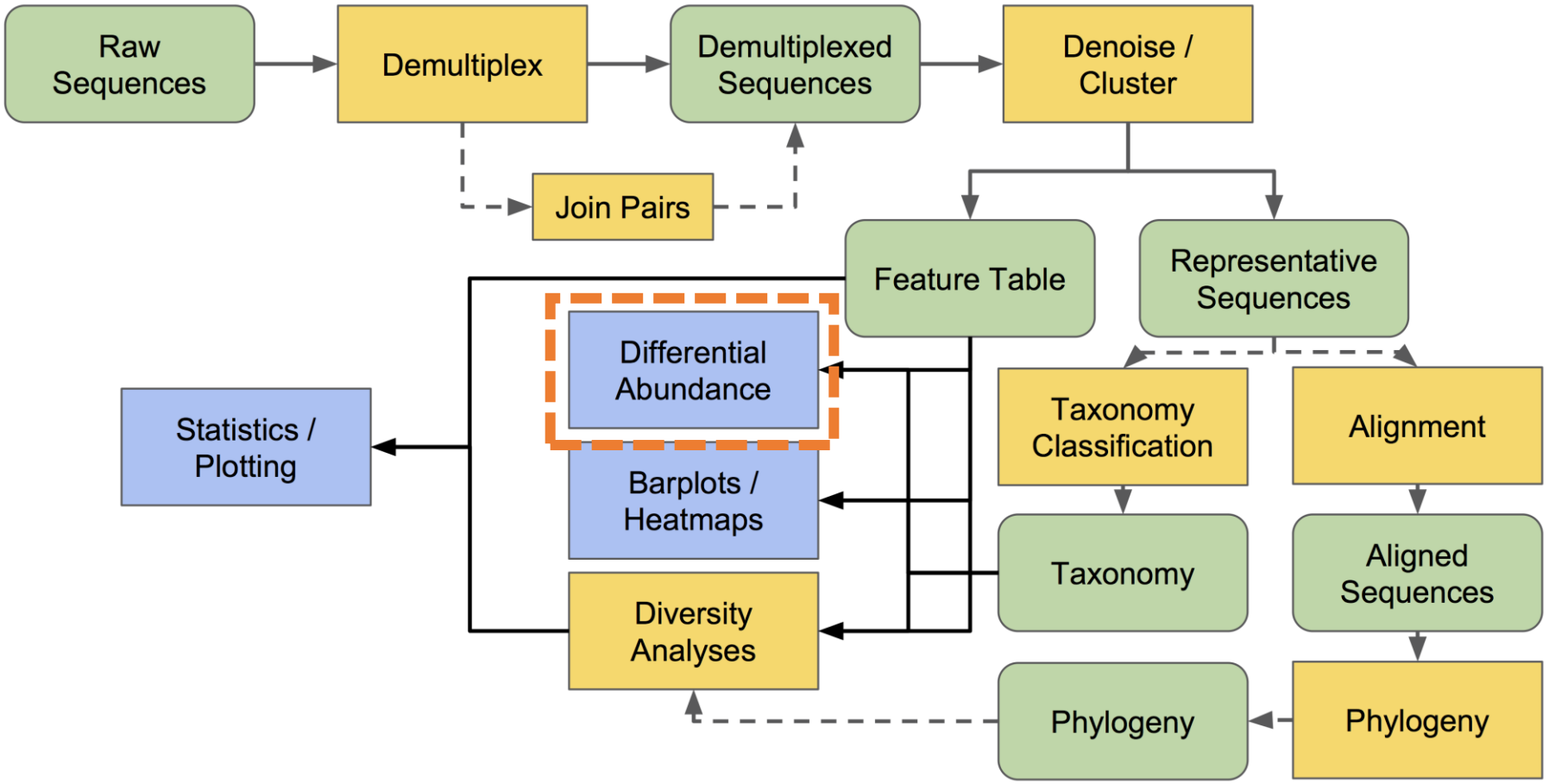
A, Richness plot stratified by anatomical location shows no difference between the number of species per sample in the 3 groups (HS lesional skin median [IQR] 72 [59-82], HS nonlesional skin 72 [63-80], and healthy controls 68 [59-78], P = .66) after rarefaction to 5000 reads. No difference was found between the 2 anatomical locations (model comparison with vs without interaction term between group and location, P = .54).

B, Shannon Diversity Index stratified by anatomical location shows the diversity of the species in each group. The index reveals an increased diversity in HS nonlesional skin (median [IQR] 3.21 [2.84-3.39]) compared with HS lesional skin and healthy controls (2.80 [2.14-3.09] and 2.52 [1.88-2.95], overall P = .005). No difference was found between the 2 anatomical locations (model comparison with vs without interaction terms between group and location, P = .91). HS indicates hidradenitis suppurativa.

Figure 4. Differential Abundance Analyses

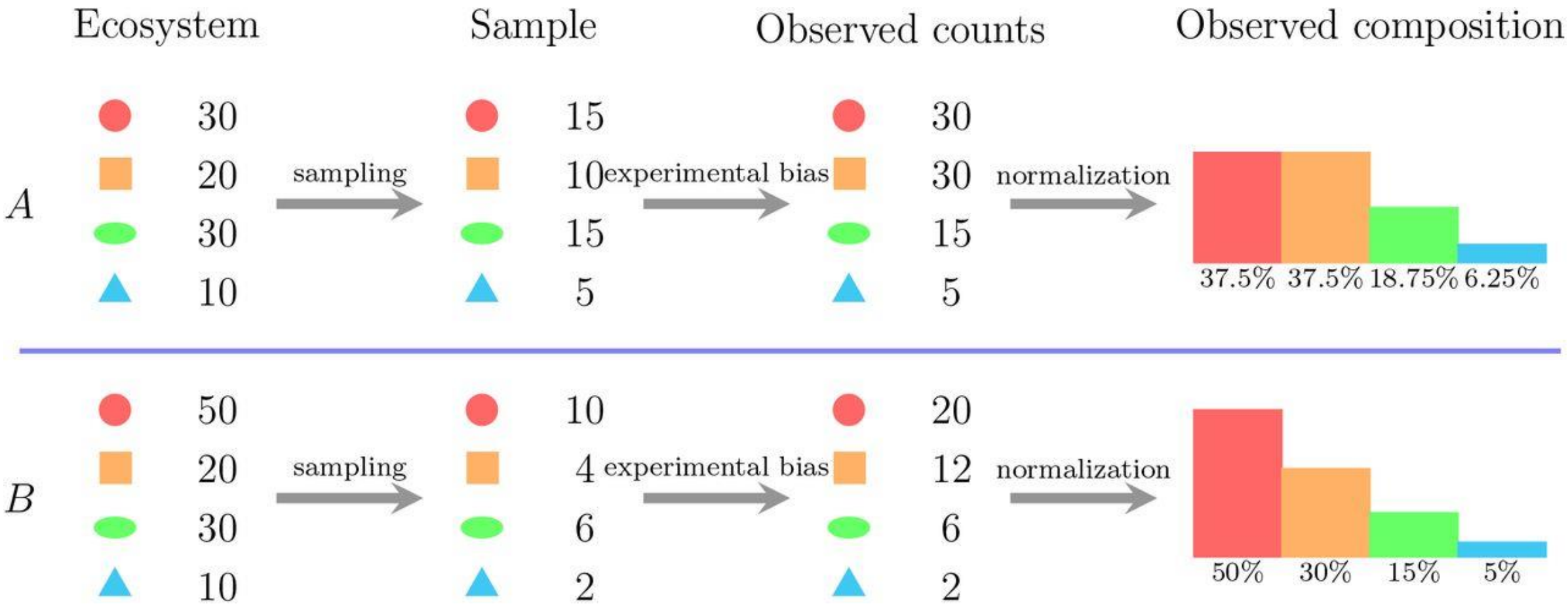


Today: Differential Abundance Analysis (DAA)



https://docs.qiime2.org/2024.5/tutorials/overview/

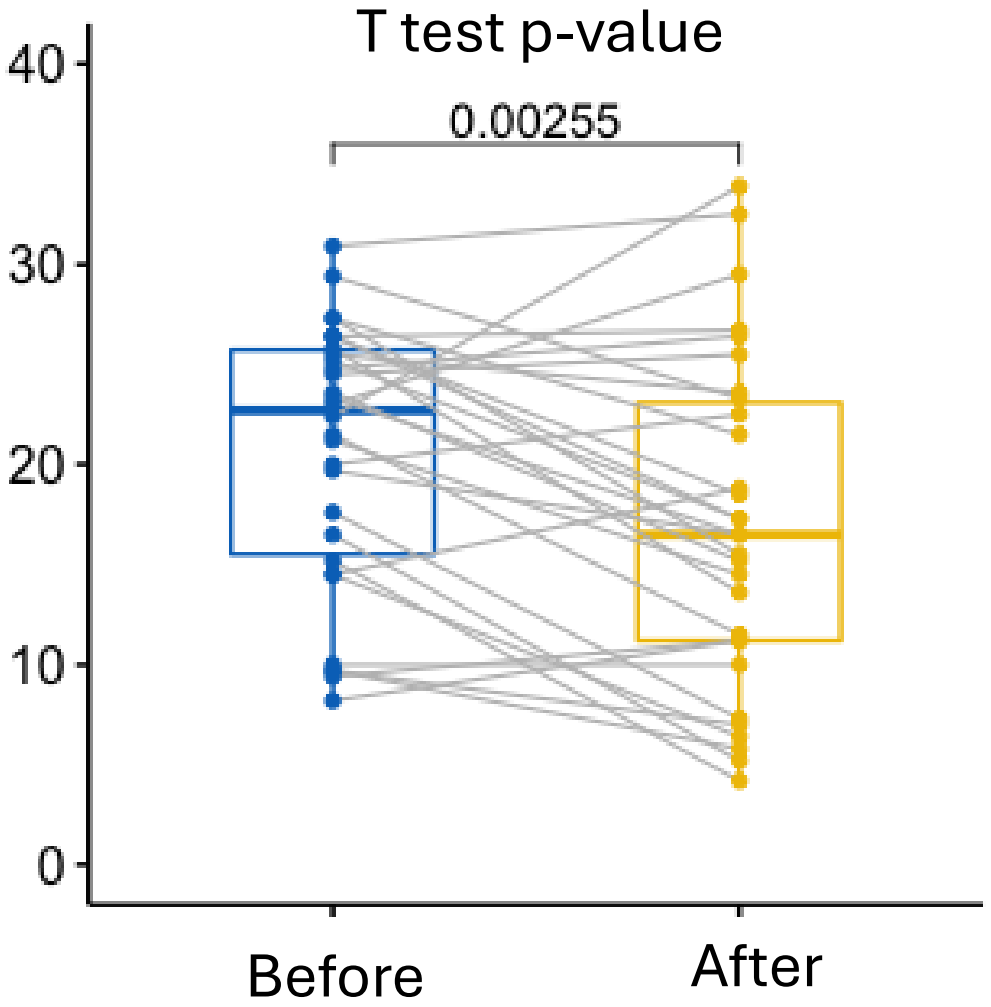
What is differential abundance (DA) analysis?



https://doi.org/10.1093/bioinformatics/btad178

How to do DAA?

Count (%) of *Propionibacterium acnes* in HS patient



treatment with drug A

Tools for DAA

Table 1 Differential abundance tools compared in this study.

Tool (version)	Input	Norm.	Trans.	Distribution	Covariates	Random effects	Hypothesis test	FDR Corr.	CoDa	Dev. For
ALDEx2 (1.18.0)	Counts	None	CLR	Dirichlet-multinomial	Yes*	No	Wilcoxon rank-sum	Yes	Yes	RNA-seq, 16S, MGS
ANCOM-II (2.1)	Counts	None	ALR	Non-parametric	Yes	Yes	Wilcoxon rank-sum	Yes	Yes	MGS
Corncob (0.1.0)	Counts	None	None	Beta-binomial	Yes	No	Wald (default)	Yes	No	16S, MGS
DESeq2 (1.26.0)	Counts	Modified RLE (default is RLE)	None	Negative binomial	Yes	No	Wald (default)	Yes	No	RNA-seq, 16S, MGS
edgeR (3.28.1)	Counts	RLE (default is TMM)	None	Negative binomial	Yes*	No	Exact	Yes	No	RNA-seq
LEFse	Rarefied Counts	TSS	None	Non-parametric	Subclass factor only	No	Kruskal–Wallis	No	No	16S, MGS
MaAsLin2 (1.0.0)	Counts	TSS	AST (default is log)	Normal (default)	Yes	Yes	Wald	Yes	No	MGS
MaAsLin2 (rare) (1.0.0)	Rarefied counts	TSS	AST (default is log)	Normal (default)	Yes	Yes	Wald	Yes	No	MGS
metagenomeSeq (1.28.2)	Counts	CSS	Log	Zero-inflated (log-) Normal	Yes	No	Moderated t	Yes	No	16S. MGS
limma voom (TMM) (3.42.2)	Counts	TMM	Log; Precision weighting	Normal (default)	Yes	Yes	Moderated t	Yes	No	RNA-seq
limma voom (TMMwsp) (3.42.2)	Counts	TMMwsp	Log; Precision weighting	Normal (default)	Yes	Yes	Moderated t	Yes	No	RNA-seq
t-test (rare)	Rarefied Counts	None	None	Normal	No	No	Welch's *t*-test	Yes	No	N/A
Wilcoxon (CLR)	CLR abundances	None	CLR	Non-parametric	No	No	Wilcoxon rank-sum	Yes	Yes	N/A
Wilcoxon (rare)	Rarefied counts	None	None	Non-parametric	No	No	Wilcoxon rank-sum	Yes	No	N/A

*The tool supports additional covariates if they are provided. ANCOM-II automatically performs ANOVA in this case, ALDEx2 requires that users select the test, and edgeR requires use of a different function (glmFit or glmQLFit instead of exactTest).
ALR additive log-ratio, *AST* arcsine square-root transformation, *CLR* centered log-ratio, *CoDa* compositional data analysis, *CSS* cumulative sum scaling, *FDR Corr.* false-discovery rate correction, *MGS* metagenomic sequencing, *RLE* relative log expression, *TMM* trimmed mean of M-values, *Trans.* transformation, *TSS* total sum scaling.

https://www.nature.com/articles/s41467-022-28034-z

Tools for DAA

Table 1 Differential abundance tools compared in this study.

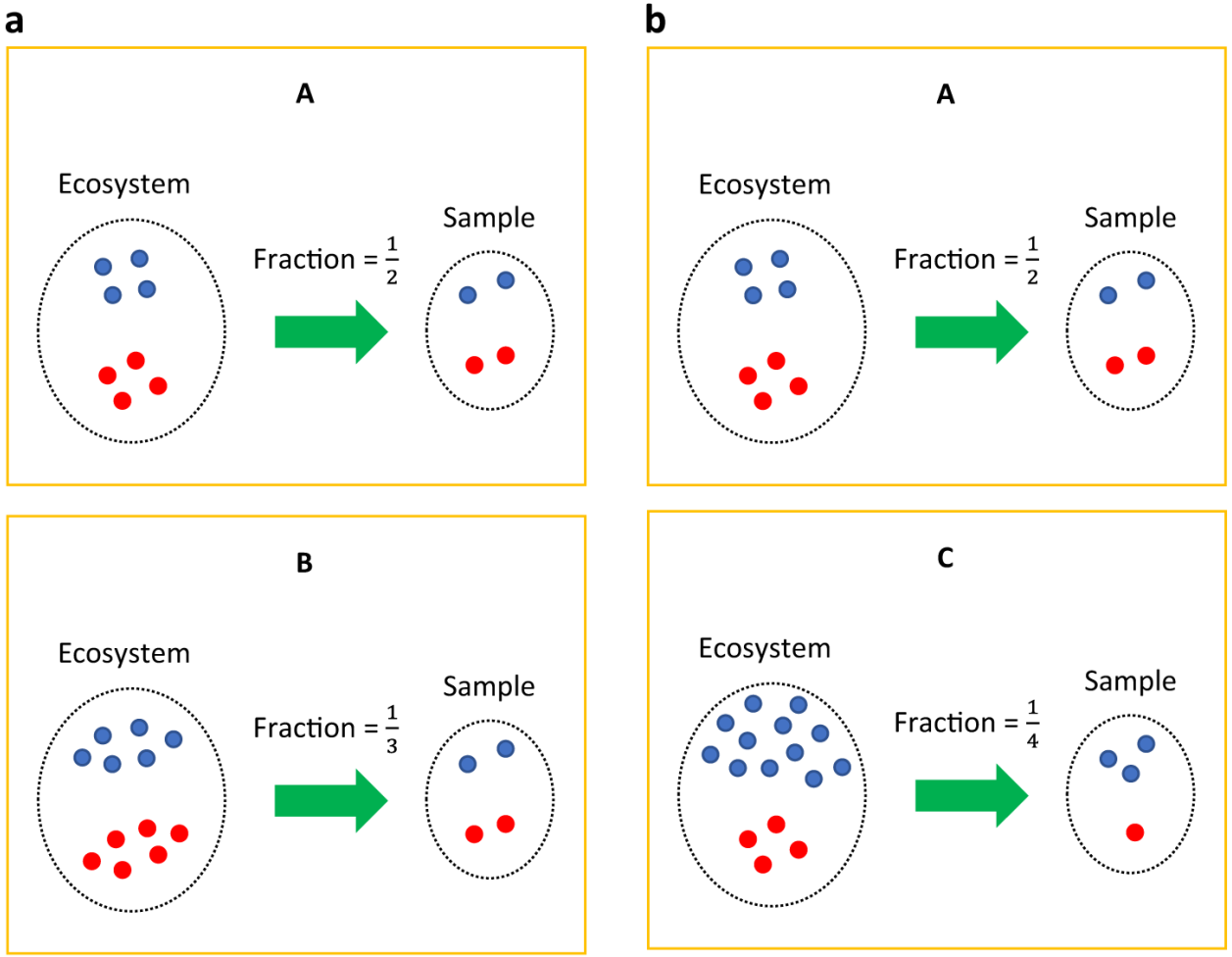
Tool (version)	Input	Norm.	Trans.	Distribution	Covariates	Random effects	Hypothesis test	FDR Corr.	CoDa	Dev. For
ALDEx2 (1.18.0)	Counts	None	CLR	Dirichlet-multinomial	Yes*	No	Wilcoxon rank-sum	Yes	Yes	RNA-seq, 16S, MGS
ANCOM-II (2.1)	Counts	None	ALR	Non-parametric	Yes	Yes	Wilcoxon rank-sum	Yes	Yes	MGS
Corncob (0.1.0)	Counts	None	None	Beta-binomial	Yes	No	Wald (default)	Yes	No	16S, MGS
DESeq2 (1.26.0)	Counts	Modified RLE (default is RLE)	None	Negative binomial	Yes	No	Wald (default)	Yes	No	RNA-seq, 16S, MGS
edgeR (3.28.1)	Counts	RLE (default is TMM)	None	Negative binomial	Yes*	No	Exact	Yes	No	RNA-seq
LEFse	Rarefied Counts	TSS	None	Non-parametric	Subclass factor only	No	Kruskal–Wallis	No	No	16S, MGS
MaAsLin2 (1.0.0)	Counts	TSS	AST (default is log)	Normal (default)	Yes	Yes	Wald	Yes	No	MGS
MaAsLin2 (rare) (1.0.0)	Rarefied counts	TSS	AST (default is log)	Normal (default)	Yes	Yes	Wald	Yes	No	MGS
metagenomeSeq (1.28.2)	Counts	CSS	Log	Zero-inflated (log-) Normal	Yes	No	Moderated t	Yes	No	16S. MGS
limma voom (TMM) (3.42.2)	Counts	TMM	Log; Precision weighting	Normal (default)	Yes	Yes	Moderated t	Yes	No	RNA-seq
limma voom (TMMwsp) (3.42.2)	Counts	TMMwsp	Log; Precision weighting	Normal (default)	Yes	Yes	Moderated t	Yes	No	RNA-seq
t-test (rare)	Rarefied Counts	None	None	Normal	No	No	Welch's *t*-test	Yes	No	N/A
Wilcoxon (CLR)	CLR abundances	None	CLR	Non-parametric	No	No	Wilcoxon rank-sum	Yes	Yes	N/A
Wilcoxon (rare)	Rarefied counts	None	None	Non-parametric	No	No	Wilcoxon rank-sum	Yes	No	N/A

*The tool supports additional covariates if they are provided. ANCOM-II automatically performs ANOVA in this case, ALDEx2 requires that users select the test, and edgeR requires use of a different function (glmFit or glmQLFit instead of exactTest).
ALR additive log-ratio, *AST* arcsine square-root transformation, *CLR* centered log-ratio, *CoDa* compositional data analysis, *CSS* cumulative sum scaling, *FDR Corr.* false-discovery rate correction, *MGS* metagenomic sequencing, *RLE* relative log expression, *TMM* trimmed mean of M-values, *Trans.* transformation, *TSS* total sum scaling.

https://www.nature.com/articles/s41467-022-28034-z

Problem: **relative abundances**

- **DA analysis should account for the bias introduced by across-sample variations in sampling fractions.**

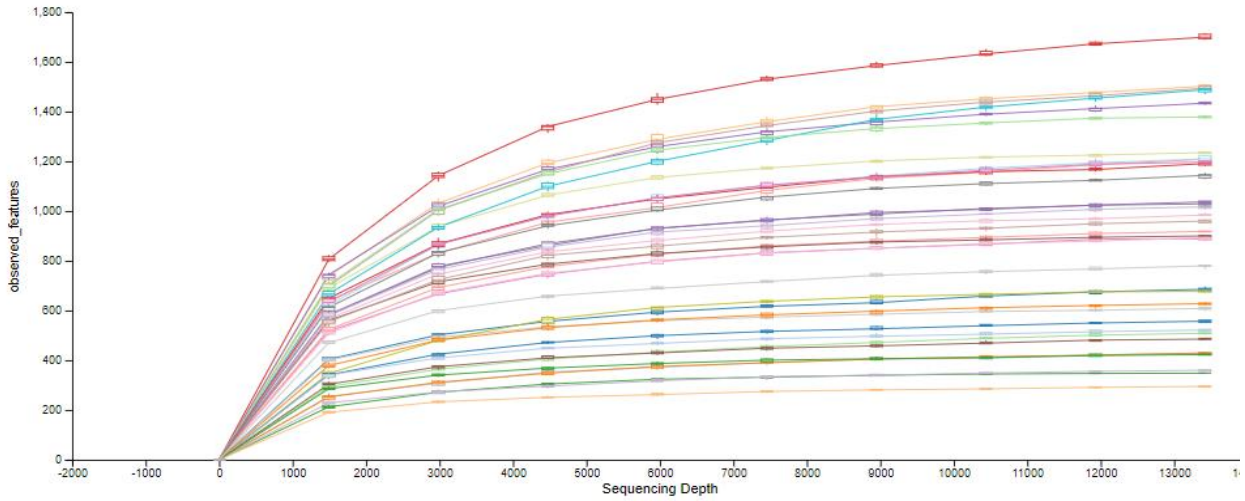


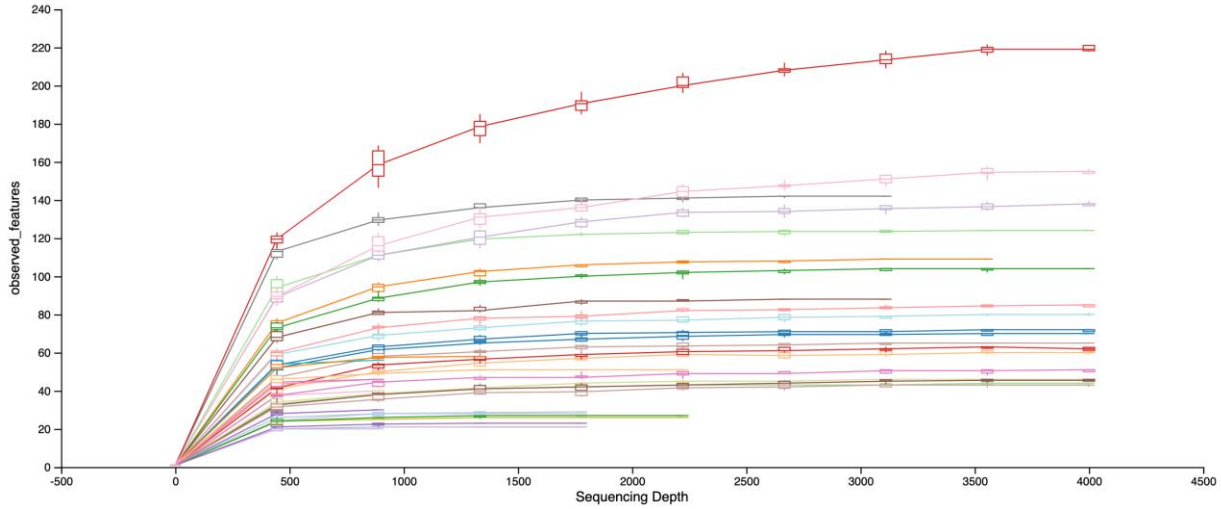
	A		B	
	Blue	Red	Blue	Red
Ecosystem	4	4	6	6
Sample	2	2	2	2

	A		C	
	Blue	Red	Blue	Red
Ecosystem	4	4	12	4
Sample	2	2	3	1

https://www.nature.com/articles/s41522-020-00160-w

Problem: **rarefied**



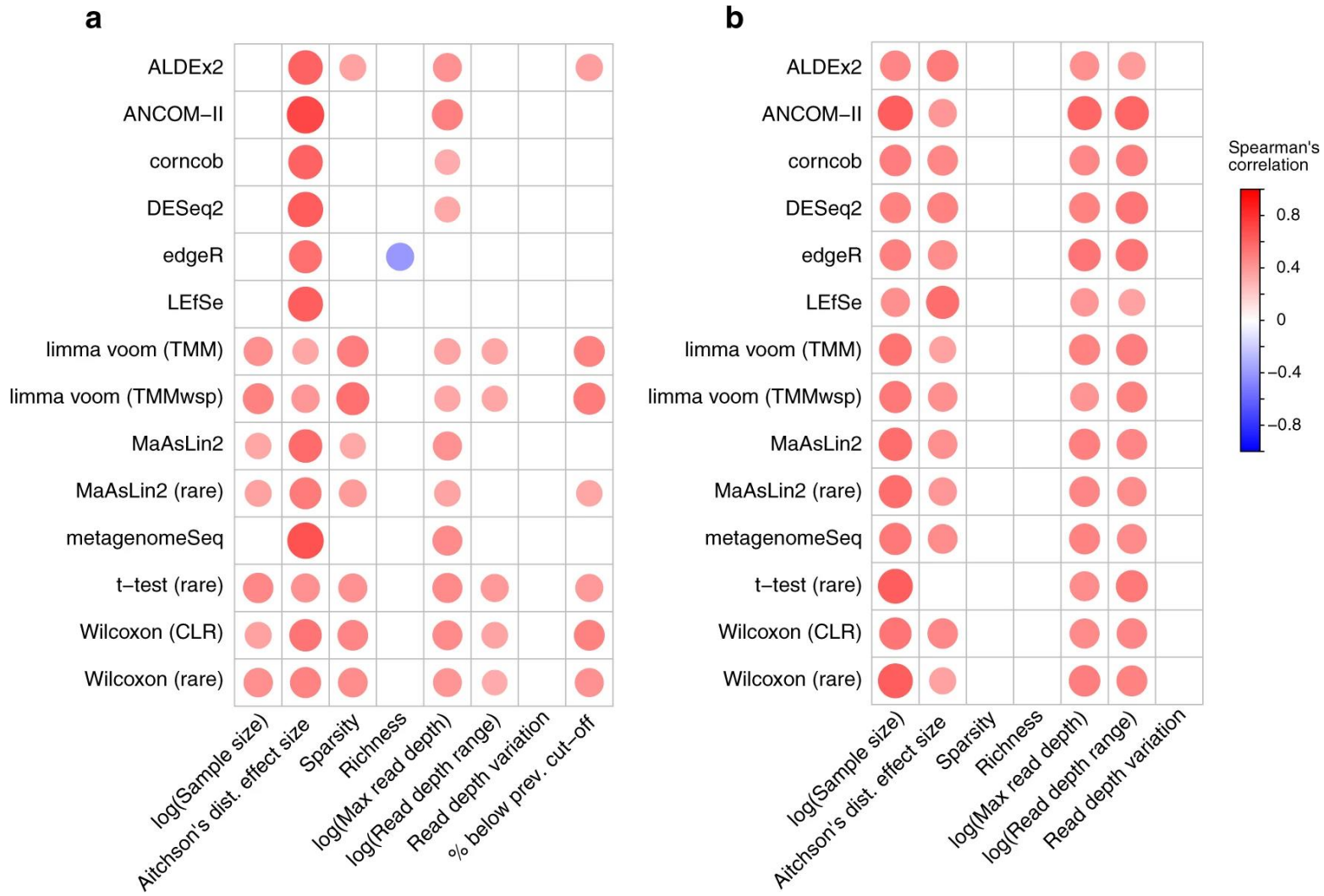


- Should read count tables be rarefied (i.e., subsampled) to correct for differing read depths across samples?

https://www.nature.com/articles/s41467-022-28034-z

Problem: **Dataset characteristics**

- **Dataset characteristics associated with percentage of significant amplicon sequence variants.**



https://www.nature.com/articles/s41522-020-00160-w

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-022-28034-z OPEN

Microbiome differential abundance methods produce different results across 38 datasets

Jacob T. Nearing[1,7✉], Gavin M. Douglas[1,7], Molly G. Hayes[2], Jocelyn MacDonald[3], Dhwani K. Desai[4], Nicole Allward[5], Casey M. A. Jones[6], Robyn J. Wright[6], Akhilesh S. Dhanani[4], André M. Comeau[4] & Morgan G. I. Langille[4,6]

s a common goal of microbiome studies. this purpose in the literature. Yet, there are g the appropriateness of using these tools of the differences between them. Here, we dance testing methods on 38 16S rRNA gene fferences in amplicon sequence variants and operational taxonomic units (ASVs) between these groups. Our findings confirm that these tools identified drastically different numbers and sets of significant ASVs, and that results depend on data pre-processing. For many tools the number of features identified correlate with aspects of the data, such as sample size, sequencing depth, and effect size of community differences. ALDEx2 and ANCOM-II produce the most consistent results across studies and agree best with the intersect of results from different approaches. Nevertheless, we recommend that researchers should use a consensus approach based on multiple differential abundance methods to help ensure robust biological interpretations.

https://www.nature.com/articles/s41467-022-28034-z

Thank you!

Projects for END Course:

1)16S (Phan Tấn Phát và Nhật Trường) 2) Shotgun (Phúc)

ELSEVIER

Water Research

Volume 246, 1 November 2023, 120700

Metaproteomics, metagenomics and 16S rRNA sequencing provide different perspectives on the aerobic granular sludge microbiome

Hugo B.C. Kleikamp [a 1], Denis Grouzdev [b], Pim Schaasberg [a], Ramon van Valderen [a], Ramon van der Zwaan [a], Roel van de Wijgaart [a], Yuemei Lin [a], Ben Abbas [a], Mario Pronk [a], Mark C.M. van Loosdrecht [a], Martin Pabst [a]

Data availability

The mass spectrometry proteomics raw data have been deposited in the ProteomeXchange consortium database with the dataset identifier PXD030677. Whole metagenome sequencing raw data are available through the NCBI Sequence Read Archive (SRA) under accession numbers SRX13522658– SRX13522660, and the 16S rRNA amplicon sequencing data under the accession numbers SRX21486087–SRX21486101. The BioProject accession number is PRJNA792132. The developed python codes for formating GTDB for the use with Diamond and QIIME are available via https://github.com/hbckleikamp/GTDB2DIAMOND and https://github.com/hbckleikamp/GTDB2QIIME.

https://www.sciencedirect.com/science/article/pii/S0043135423011405

Projects for END Course:

Microbiome Sequencing and Analysis (Project 3: Nguyễn Hữu An and Trần Thị Vũ)

Biopsy specimens were analyzed at the Department of Microbiology and Infection Control, Statens Serum Institut, Copenhagen, Denmark. DNA was extracted using a kit (QIAamp DNA Mini Kit; Qiagen) according to the manufacturer's instructions for tissues. For each batch of DNA extraction, a "negative" control was included containing buffers but no sample material for downstream analysis. DNA was amplified using a 2-step polymerase chain reaction using custom 341F/806R primers targeting the V3-V4 16S regions, as well as 3 primer sets targeting the hypervariable regions V3-V4 of the *18SrDNA* gene, and amplicons were sequenced on a desktop sequencer (MiSeq; Illumina, Inc) using the v2 reagent kit. For details concerning primer design and library preparation, see the eAppendix in the Supplement. Sequence data are available at the European Nucleotide Archive (accession number **PRJEB15266**).

https://jamanetwork.com/journals/jamadermatology/fullarticle/2628754

Projects for END Course:

4)16S tôm/cá (Ng T Anh Thư và Khải)

PeerJ JOURNALS PUBLISH COMMUNITIES

‹ AQUATIC BIOLOGY

Microbial signature profiles of *Penaeus vannamei* larvae in low-survival hatchery tanks affected by vibriosis

PeerJ Life & Environment

Research article | Aquaculture, Fisheries and Fish Science | Bioinformatics | Microbiology | Molecular Biology | Histology

Related research

Share

Guillermo Reyes[1], Betsy Andrade[1], Irma Betancourt[1], Fanny Panchana[1], Ramiro Solórzano[1], Cristhian Preciado[1], Lita Sorroza[2], Luis E. Trujillo[3], Bonny Bayot[1,4]

Published September 1, 2023

(A260/280 = 1.8–2.0) were examined through a NanoDrop One Microvolume UV-Vis Scanning Spectral Spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). The gDNA of all 45 samples was then submitted to Novogene Incorporation (Beijing, China) for amplification of the V3–V4 region of the *16S rRNA* gene using the 341F/806R primers (Takahashi et al., 2014), amplicon library construction, and paired-end sequencing with the Illumina NovaSeq 6000 P250 platform. The DNA sequences of the 20 samples from the five tanks affected by zoea 2 syndrome were deposited at the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under accession number PRJNA870476. The sequences of the 25 samples from the seven tanks affected by AHPND are available under accession number PRJNA800805 (Reyes et al., 2022).

https://peerj.com/articles/15795/

https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA870476

Projects for END Course:

5) Shotgun bệnh (Ng Kỳ Phát, Trương Anh Tú, Trần Quang Minh và cô Diệp)

Research Paper

Preterm infants at low risk for early-onset sepsis differ in early fecal microbiome assembly

Sagori Mukhopadhyay, Jung-Jin Lee, Erica Hartman, Emily Woodford, Miren B. Dhudasia, Lisa M. Mattei,

...show all

Article: 2154091 | Received 06 Jul 2022, Accepted 28 Nov 2022, Published online: 06 Dec 2022

Cite this article https://doi.org/10.1080/19490976.2022.2154091 Check for updates

The shotgun metagenomic sequencing data supporting the conclusions of this article is deposited at the NCBI Sequence Read Archive (PRJNA872399) and can be accessed at the URL https://www.ncbi.nlm.nih.gov/bioproject/PRJNA872399.

https://www.tandfonline.com/doi/full/10.1080/19490976.2022.2154091

Projects for END Course:

6) 16S bệnh (Hoàng Anh, Hoàng Sơn và Ngô Đại Phú)

Các bạn điền vào giúp, cảm ơn các bạn!