Metagenomics and the microbiome

London School of Hygiene and Tropical Medicine

Microbiome

Background

The **microbiome** is the collection of genetic material of the microbial flora in an environment

E.g., bacteria, fungi, viruses, and their genes
Including, those that naturally live on our bodies and inside humans

Studying the different types and functions of microbes in their natural environment without growing them in a lab

The community is accessed by sequencing of linked samples (e.g., faecal -> gut microbiota)

The relative abundance of different microbial genetic signatures is assessed

Two main sequencing approaches:

Two main sequencing approaches:

Metagenomics
(Shotgun sequencing)

- Attempts to sequence everything in a sample
- Untargeted sequencing approach, reads all genomic DNA in a sample
- Generates **millions of reads** (more than most microbial projects)
- Its sensitivity makes it an attractive choice for clinical use.

Targeted sequencing
(Amplicon analysis)

- Amplify + sequence a marker gene
- Just one specific region of DNA (e.g. **16S rRNA**)
- Might recover diversity well, but biased depending on region amplified
- Can be used to identify samples for a metagenomic approach

Two main sequencing approaches:

SHOTGUN "Metagenomics"

PROS

- No need for specific prior knowledge
- Rich data → Greater potential insights (e.g., function)

CONS

- Protocols and data cleaning may introduce bias
- Complex analysis due to data diversity and volume

Amplicon / targeted sequencing

PROS

- Less prone to contamination
- Simpler QC (easier to spot contaminants)

CONS

- Requires understanding of the microbial community for primer design
- Are we capturing enough variation?

16S microbiome analysis

Understanding prokaryotic diversity

- Common approach:
  Use the gene coding for the prokaryotic **16S ribosomal RNA**

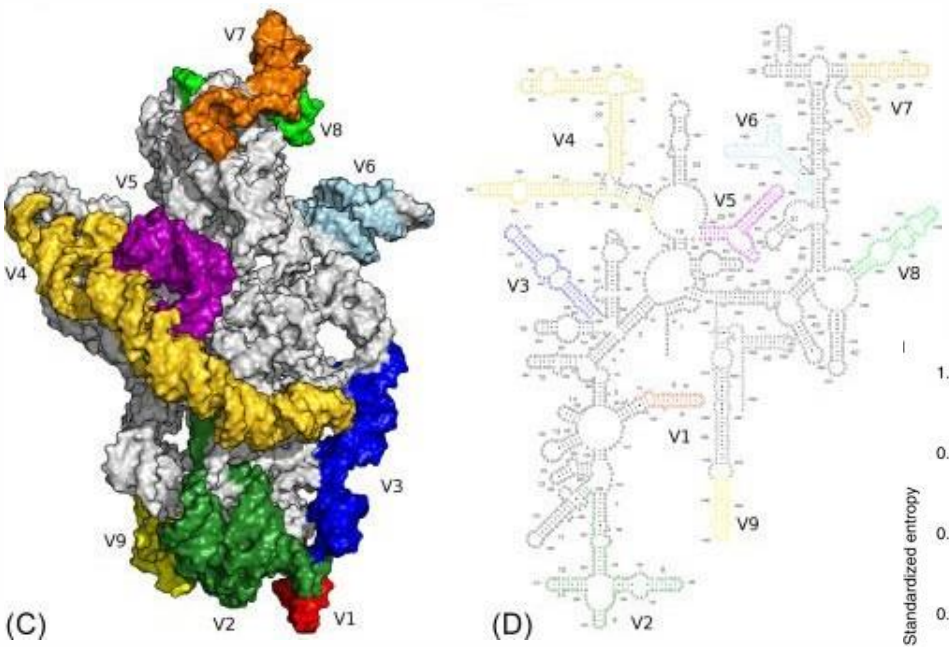
  - ***The Universal Marker:***
    Present in all bacteria and archaea.

  - ***Structural composition:***
    Split in conserved + (hyper-)variable regions (V1-V9) → ideal for priming

  - ***Research Significance:***
    Extensively studied, provides a wealth of comparison data

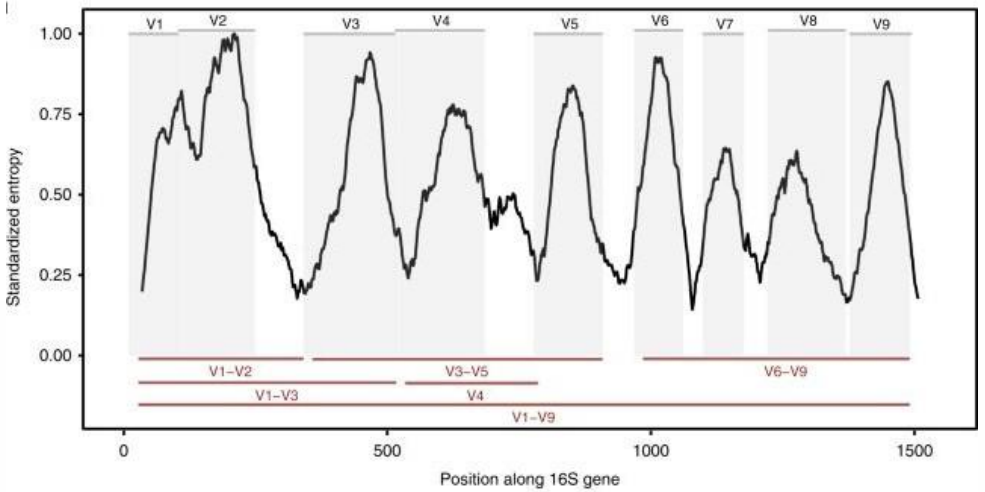
  - ***Limitations:***
    - Gives only information about the relative abundance of individual taxa and not metabolic functionality
    - Some species have the same sequence in some variable regions and / or multiple copies of the 16S gene

The 16S gene



https://doi.org/10.1016/B978-0-08-102268-9.00005-7

V1-V9



https://doi.org/10.1038/s41467-019-13036-1

Operational taxonomic Units (OTUs)

A collection of 16S rRNA sequences that have a certain percentage of sequence divergence

Clustering based on a user defined identity threshold (e.g. ~97%)

Approaches to defining OTUs:

- ***de novo:*** Clusters sequences without using a reference genome
- **open reference:** Aligns with a database + include non aligned clusters
- **closed reference:** Uses a predefined database for clustering

Challenges and Limitations:

- May merge closely related species, losing detail
- Hard to include new data / compare studies
- Small species differences can be as minor as a single nucleotide change
- Distinguishing true biological variation from sequencing errors remains a key challenge.

Amplicon sequence variants (ASVs)

The ASV method quantifies each unique sequence by its read frequency, employing an error model to discern true sequences from sequencing errors.

This generates a probability score for each sequence, assessing whether it's an error artifact or a valid genetic variant.

Advantages:

- Left with only sequences with high statistical confidence
- Allows the addition of new sample data
- No reference bias

Challenges:

- May overlook low-abundance species
- Requires substantial computational resources

OTU vs ASV

Which one is better?

OTU	ASV
Not easily shared across studies	Easily shared across studies
Prone to reference bias	Independent of reference until classification
Averages sequences into a consensus	Captures individual sequences exactly
May group diverse species	Specific to a shared sequence
Subject to chimeras	Subject to chimeras
Complex chimera identification	Simplified chimera identification

ASV approach is more widely used these days

Next Steps: Analysing the Counts Table

After denoising / clustering the counts table is acquired to show ASV/OTU frequencies

Analysis Objectives:

- ***Alpha diversity:***
  Assess species diversity within each sample/group
- ***Beta diversity:***
  Compare species diversity between different samples
- ***Rarefaction:***
  Discover any potential diversity missed
- ***Taxonomic Assignment:***
  Classify species using phylogenetics
- ***Differential Abundance Analysis:***
  Identify statistically significant variations in species across conditions

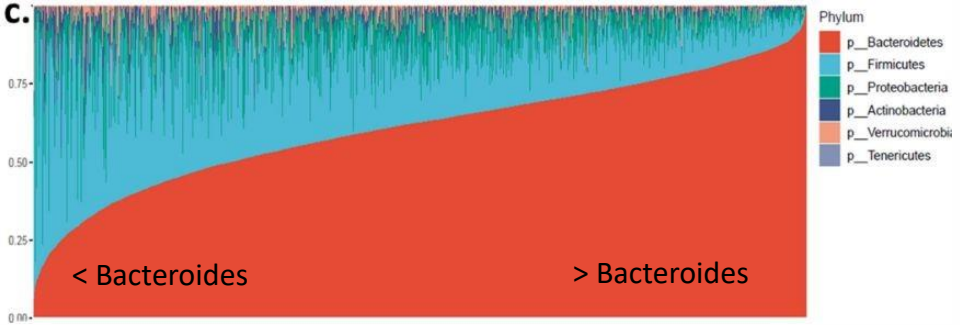


Figure: Sample Distribution by Relative Abundance ordered by Bacteroides abundance

https://doi.org/10.1101/2020.11.27.401125

Composition Analysis: A Cautionary note

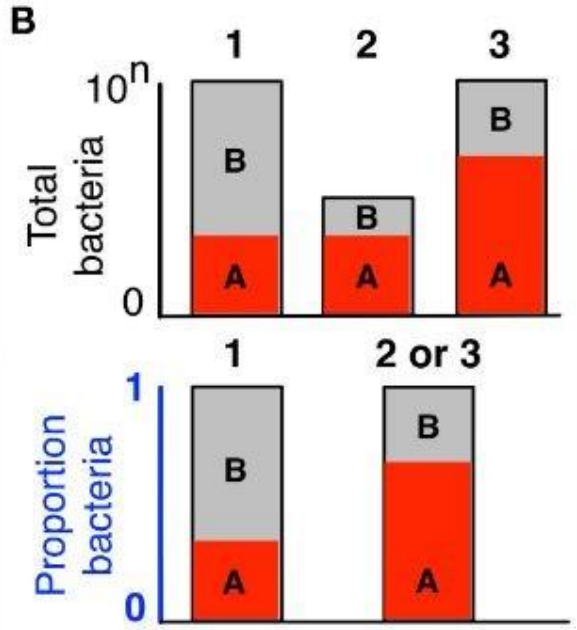
Microbiome Datasets Are Compositional: And This Is Not Optional

Gregory B. Gloor[1], Jean M. Macklaim[1], Vera Pawlowsky-Glahn[2] and Juan J. Egozcue[3]*

[1] Department of Biochemistry, University of Western Ontario, London, ON, Canada, [2] Departments of Computer Science, Applied Mathematics, and Statistics, Universitat de Girona, Girona, Spain, [3] Department of Applied Mathematics, Universitat Politècnica de Catalunya, Barcelona, Spain

- Data reflects relative, not absolute, abundancies
- Standard statistical methods may yield misleading results with compositional data
- Employ specialised methods (e.g. ANCOM)
- Consider data transformation techniques (e.g. centre-log ration) for clarity

Samples 1 and 2 appear to have the same number of A bacteria…



Using proportional data shows Sample 2 is more like sample 3

Practical

- ***Quality control:***

  Ensure the reliability of multiple sequences

- ***Denoising:***

  Refine the data removing any noise

- ***Classification:***

  Assign taxa based on phylogenetics

- ***Rarefaction:***

  Discover any potential diversity that may have been missed

- ***Visualising Diversity:***

  Graphically represent alpha and beta diversity

We will be using Qiime2, but other tools exist such as Mothur, USEARCH, STAMP

References

- Cho, Ilseung, and Martin J. Blaser. "The human microbiome: at the interface of health and disease." *Nature Reviews Genetics* 13.4 (2012): 260-270.
- Oulas, A., Pavloudi, C., Polymenakou, P., Pavlopoulos, G. A., Papanikolaou, N., Kotoulas, G., … Iliopoulos, I. (2015). Metagenomics: Tools and Insights for Analyzing Next-Generation Sequencing Data Derived from Biodiversity Studies. *Bioinformatics and Biology Insights*, *9*, 75–88. JOUR. http://doi.org/10.4137/BBI.S12462
- Jovel, J., Patterson, J., Wang, W., Hotte, N., O'Keefe, S., Mitchel, T., … Wong, G. K.-S. (2016). Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics. *Frontiers in Microbiology*, *7*, 459. JOUR. http://doi.org/10.3389/fmicb.2016.00459
- Nayfach, S., & Pollard, K. S. (2016). Toward Accurate and Quantitative Comparative Metagenomics. *Cell*, *166*(5), 1103–1116. http://doi.org/10.1016/j.cell.2016.08.007
- Gloor, Gregory B., et al. "Microbiome datasets are compositional: and this is not optional." *Frontiers in microbiology* 8 (2017): 2224.
- Ramazzotti, Matteo, and Giovanni Bacci. "16S rRNA-based taxonomy profiling in the metagenomics era." *Metagenomics*. Academic Press, 2018. 103-119.
- Johnson, Jethro S., et al. "Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis." *Nature communications* 10.1 (2019): 1-11.
- Gacesa, Ranko, et al. "The Dutch Microbiome Project defines factors that shape the healthy gut microbiome." *BioRxiv* (2020).
- http://www.illumina.com/areas-of-interest/microbiology/microbial-sequencing-methods/shotgun-metagenomic-sequencing.h tml

Videos

- Good general overview: https://www.youtube.com/watch?v=6564K4-_DBI&list=PLOPiWVjg6aTzsA53N19YqJQeZpSCH9QPc&index=2
- ASVs vs OTUs: https://www.zymoresearch.com/blogs/blog/microbiome-informatics-otu-vs-asv