Metagenomics and the microbiome

London School of Hygiene and Tropical Medicine

- ***In situ***, **culture-free** genomic characterization of the **taxonomic and functional profiles** of a **microbial community**.

- Identifies and quantifies microbial taxa and/or genes, to know **"who" is there** and **what functions** they can perform.

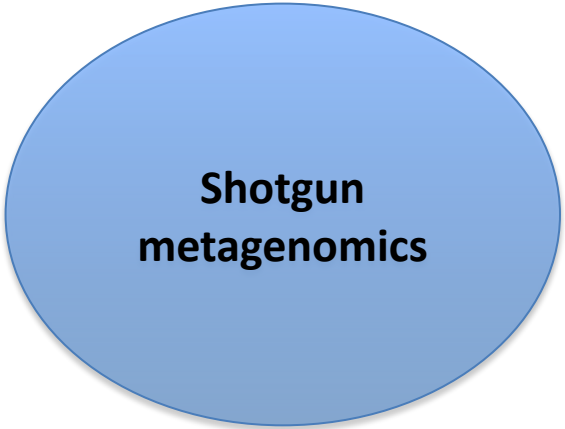
"Shotgun" vs targeted sequencing

- Sequence everything
- Sometimes also called "metagenomics"
- Generates **millions of reads** (more than most microbial / parasite projects)

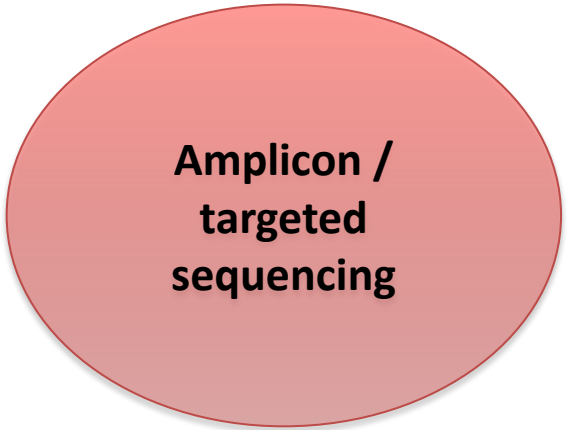
- Scale of data challenging
- experimental protocols (DNA extraction, library preparation, etc.) + data cleaning introduce bias

- Amplify + sequence a marker gene (e.g. 16S rRNA)
- Might recover diversity fairly well but biased depending on region amplified

- no direct information on metabolic functionality of ecosystem



- Less *a priori* knowledge before processing sample: we take the sample and we sequence it.
- Rich data → more potential insight (functionality etc.)
- Analysis more complex due to diversity and size of the data.
- More expensive to sequence to relevant depth



- Extra amplification step
- More *a priori* knowledge about the community needed (primer selection)
- Simpler QC (easier to spot obvious contaminants)
- But: Are we capturing enough variation (strain variability)?

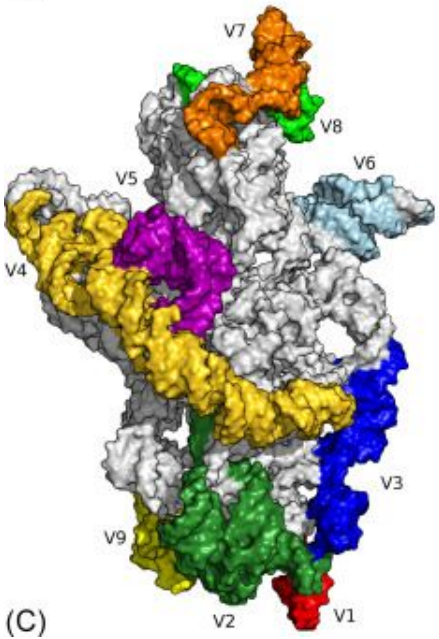
16S microbiome analysis

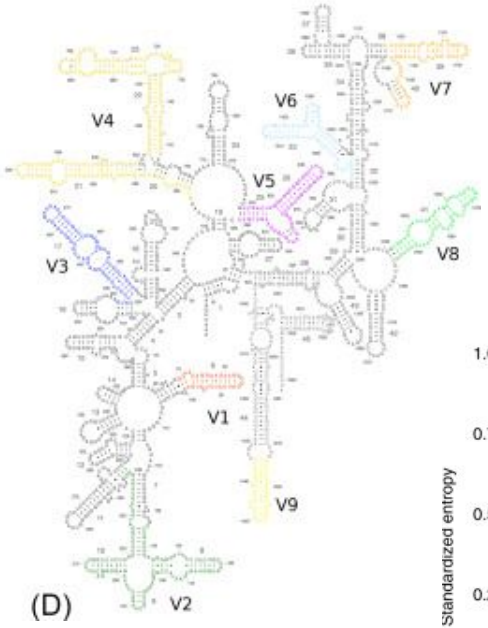
- The **microbiome** is the collection of genetic material of the microbial flora in an environment (e.g. on or in a human host)

- Most microbiome studies use the gene coding for the prokaryotic **16S ribosomal RNA**
    - Has a structural role as a scaffold defining the positions of proteins in the small ribosomal unit.
    - Present in all bacteria and archaea.
    - Split in conserved + (hyper-)variable regions (V1-V9) → ideal for priming

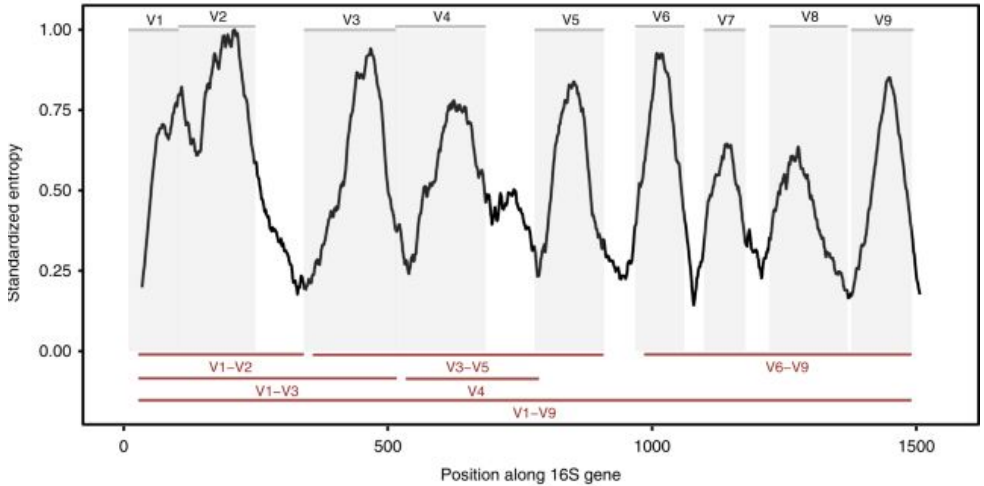
- Gives only information about the relative abundance of individual taxa and not metabolic functionality etc.
- some species have the same sequence in some variable regions and / or multiple copies of the 16S gene

The 16S gene





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



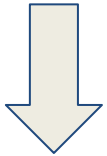
https://doi.org/10.1038/s41467-019-13036-1

16S methodology – OTUs vs ASVs

Difference between species can be as small as a single nucleotide
→ main challenge to distinguish biological variation from sequencing errors

Clustering based on identity threshold (usually 97%):

- *de novo*
- open reference
- closed reference



taking more time

→ generates OTUs (operational taxonomic units)

- ignores details + combines closely related species
- hard to include new data / compare studies

Denoising:
Fits error model and estimates probability of read being original or due to sequencing error

→ generates ASVs (amplicon sequence variants)

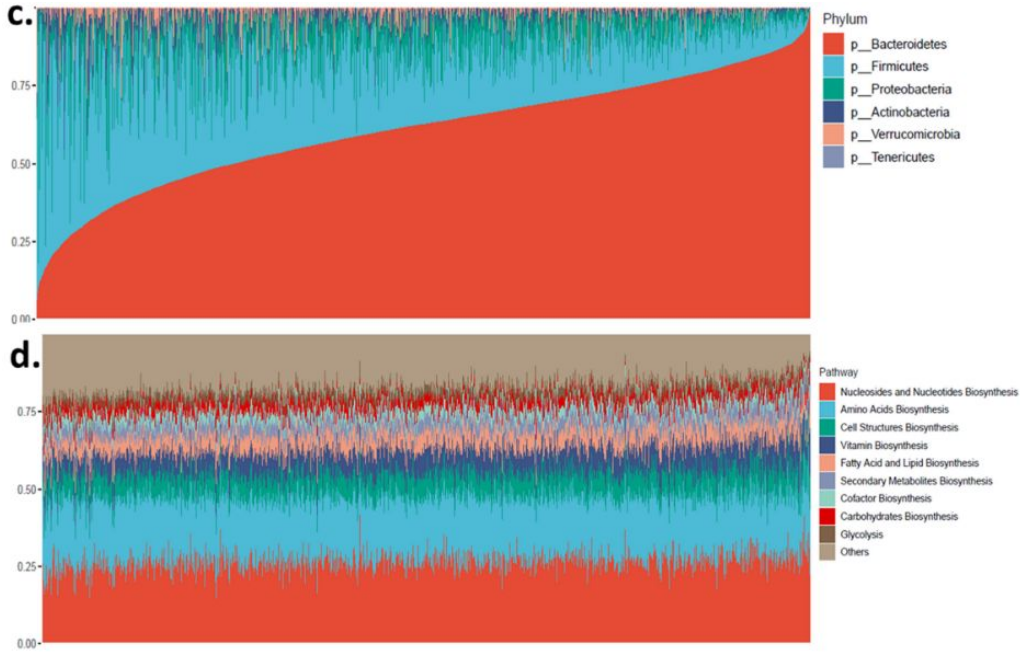
- might not recognise some low-abundance species
- can be computationally expensive

We got a counts table – what now?

After denoising / clustering we know how often each ASV / OTU appeared in each sample → “counts table”.

Common aims of downstream analysis:

- alpha (within samples / groups) and beta (between samples / groups) diversity
- rarefaction
- taxonomic assignment + phylogenetics
- differential abundance + regression
- clustering + enterotypes (debated)
- infer functionality (debated)



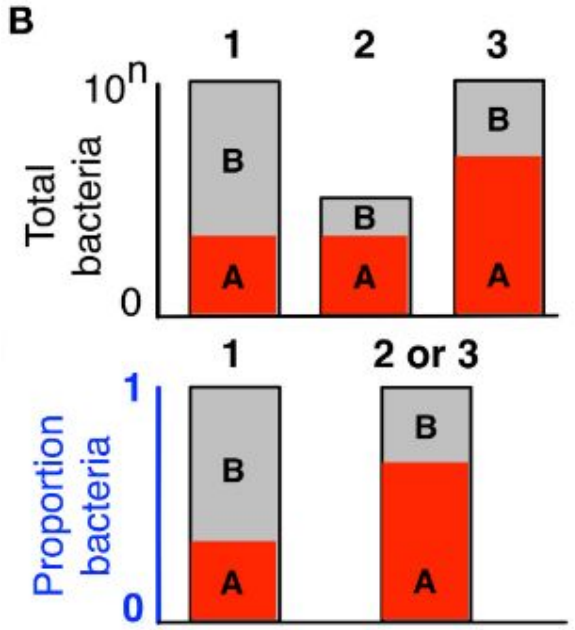
https://doi.org/10.1101/2020.11.27.401125

Compositionality – a word of caution

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



Most statistical assumptions not satisfied in compositional data → many default methods will give spurious results

→ use specialised methods (e.g. ANCOM)

→ transform data (e.g. center-log ratio)

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.html

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