Introducing Microbiome Bioinformatics

Part 1.

Aims of these sessions (1)

- Overview of types of **microbiome analysis**
  - with particular regard to **sequence informatics concepts**
- “Top down” – putting analysis tools and resources in context:
- How features of those experimental platforms dictate the bioinformatics approaches
- Why the nature of the data gives rise to the-
  - Databases
  - Software
  - Algorithms
- - that are commonly used

Aims (2)

- Explore pros and cons of different approaches
- Different sequencing ‘omics
  - **16S** (and analogous) “barcoding”
  - “Shotgun” metagenomics
  - Metatranscriptomics
- Problems and possible solutions
  - Consistency
  - Errors and bias
- Computing environments, software and skills

Aims (3)

- Main audience:
  - those who are doing the analyses
  - and/or **planning the experiments**
  - plus anyone else interested ☺
- No highly detailed technicalities
  - No instructions on how to run particular programs
- I'll have been successful if…
  - You understand why you are using the bioinformatics approaches you use
  - What's good and bad about them
  - And that alternatives may be available!

Future talks

- At least 2 (probably 3) sessions to cover what I would like to
- Beyond that – if there is demand –
  - can progress to more technical talks
  - especially about 16S analysis (probably)
  - increasingly metagenomics in GHFS research
- Informal and flexible
  - Please interrupt and ask questions
  - Suggestions for topics for further focus

Part 1

- Informatics-relevant aspects of:
  - Biology
  - Microbes, genes and genomics
  - Scientific aims
  - Sequencing Platforms
- “Biological and Experimental Stuff that a microbiome bioinformatician needs to know”

clker.com

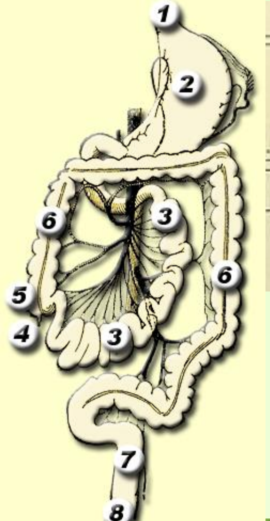
- **Who is in there?**
  - In what amounts?

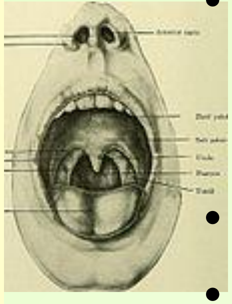
Analysis of **marker genes** ("barcodes")
e.g. for **prokaryotes**: 16S rRNA gene
"**16S-barcoding**"

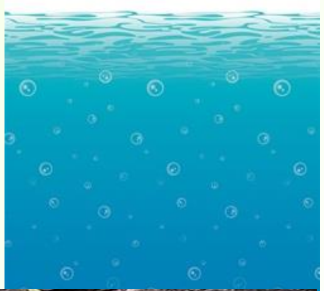
- Who is in there….
- **…and what are they doing?**

Shotgun Metagenomics - what can they do?
Metatranscriptomics - what are they doing?
Proteomics
Metabolomics

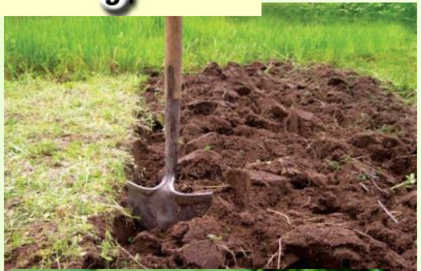
Studied environments

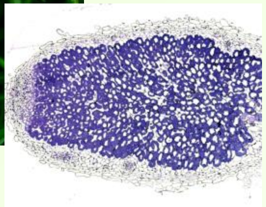


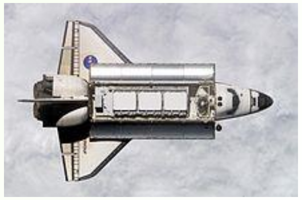




- Human organs and tracts
  - Especially gut; mouth, nose, skin, genitals, everywhere…
- Animal intestinal tracts, organs…
- Aquatic environments of all kinds
- Soil, plant and plant-related
  - Bulk soil, rhizosphere, mycorrhizal, leaves
- Biofilms - of many kinds
  - In civic and industrial infrastructure, clinical
- Debris you scrape off your number-plate after a long road-trip
- Rock samples sent into space and back…





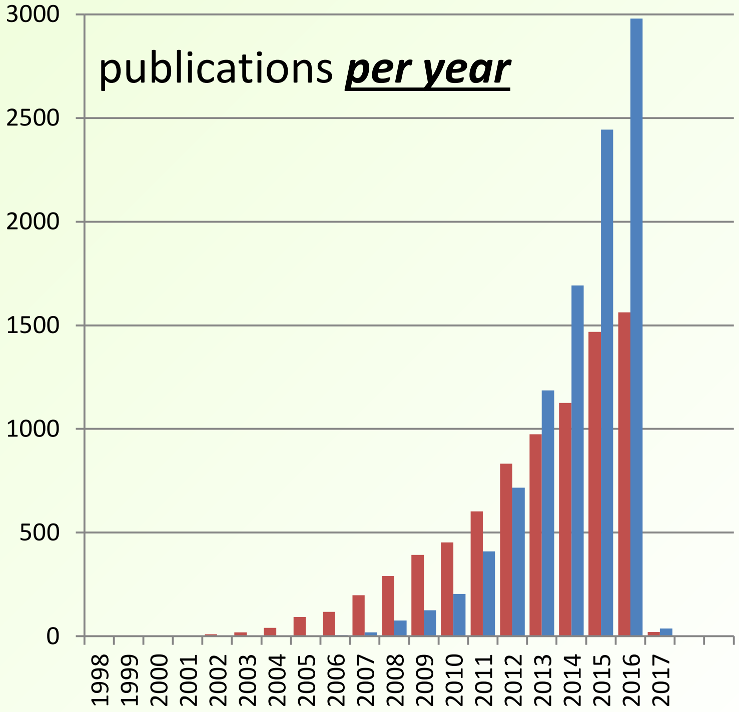


Some scientometrics

As **currently indexed** (Jan 2017)
In WoS core collection:
keyword search by ***Topic***

metagenomic* OR metagenome* OR microbiome* OR metatranscriptomic* OR metatranscriptome* AND…

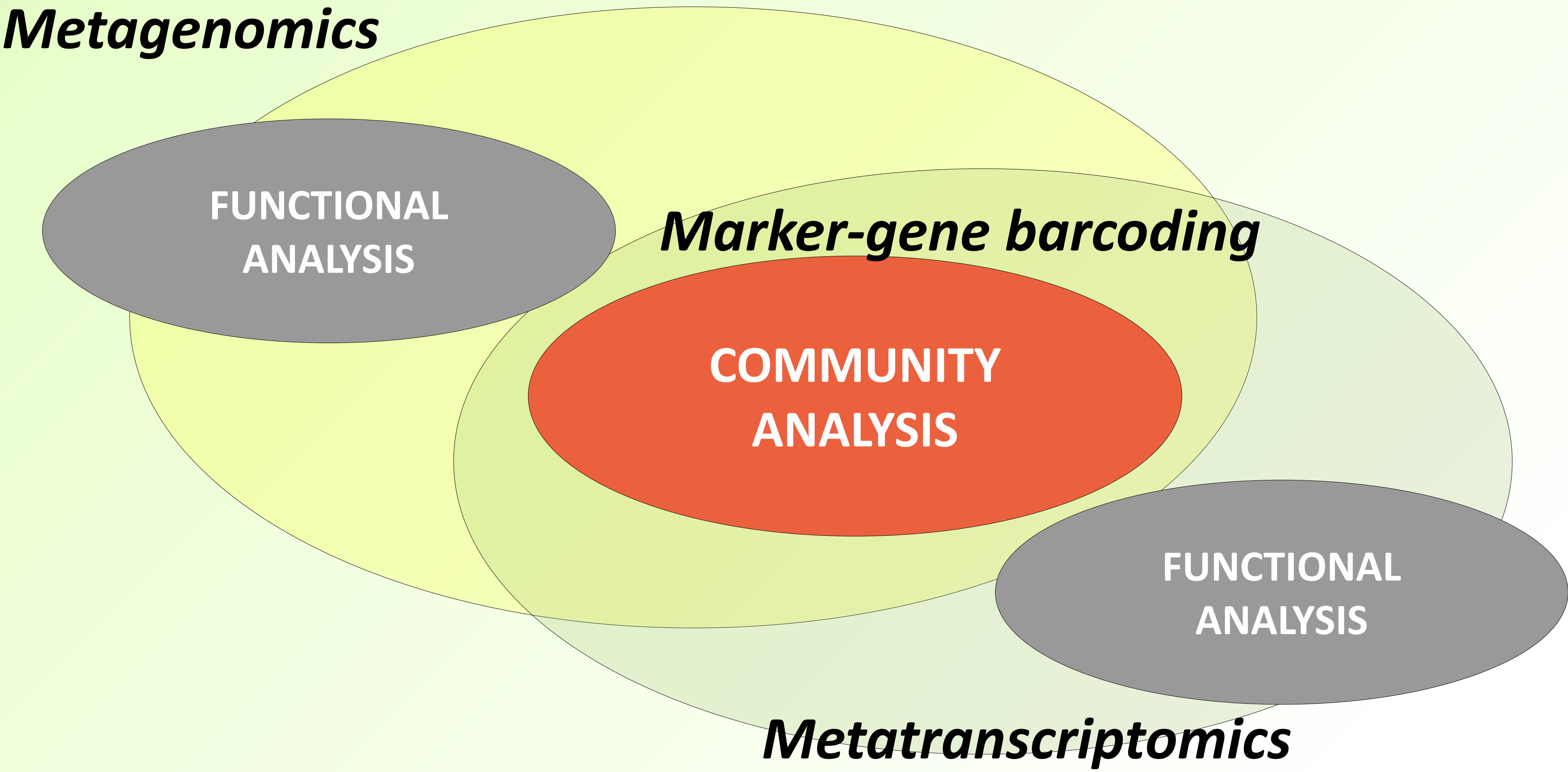
(all)	17,173
gut, GIT, gastrointestinal	6,024
plant , soil , rhizosphere , rhizoplane , phyllosphere	2,982
aquatic , marine , ocean , lake	1,985
virus , phage , virome	1,749
bioinformatic* , computation*	1,127



Estimated 5-10% of these focus on bioinformatics (software/ databases / practices)

Analysing Microbiomes

- Aims of the analysis –
  - **communities, functions**
- What do you sequence, and why?
- How do you process the sequence reads?
- Which software do you need?
- Which databases?
- What can you conclude?


Metagenomics
FUNCTIONAL ANALYSIS
Marker-gene barcoding
COMMUNITY ANALYSIS
FUNCTIONAL ANALYSIS
Metatranscriptomics

Aims of Community Analysis

- **Diversity studies**
  - How similar are the members of the community?
    - Within samples
    - Between samples
  - We may not necessarily care about identifying the community members; just how different they are
- Richness and diversity are just about the only part of microbiome informatics that you **can** do:
  - **without needing any reference databases**
- In practice, it is normal to use reference data as well

Aims of Community Analysis (2)

- **Identifying** the members of the community
  - Which species, genera, classes, phyla etc are present?
  - Differences between samples (treatment versus control)
  - Clearly requires using reference data:
    - Defined taxonomic systems
    - More on this in a later session
  - Can we find evidence of biological significance of particular groups of organisms?

Who can we identify?

- Viruses
- Prokaryotes
  - Bacteria
  - Archaea
- Eukaryotes
  - Fungi
  - Oomycetes
  - Ciliates
  - Flagellates
  - Metazoans
    - Nematodes
    - Insects
    - Etc…
  - Plants
  - Etc…
- … *whatever is in the reference databases*

- A lot of metagenomics is **prokaryote-centric**
  - Many environments are heavily populated by bacteria
  - High cell count = High copy number of DNA sequences
  - Genomes have high gene-density
  - Very large number of **reference genome sequences**
- **Useful quantities of eukaryote DNA can be recovered**
- Also true of mRNA/rRNA in metatranscriptomics
  - e.g. 3%-20% of sequence recovered from **rhizosphere**
- **The Virome:**
  - **RNA viruses, DNA viruses**
  - **Single-stranded, double-stranded**

Aims of functional analysis: **What** can we identify?

- All kinds of DNA sequence
  - Genes
    - Coding
    - Non-coding
      - Small RNA etc
  - Intergenic sequence
- … *whatever is in the reference databases*

- Metatranscriptomics:
  - All kinds of transcripts
  - Potentially some challenges with sequencing some kinds of transcripts
- *Again – reference data required*

Is there anything else?

- Can we **shotgun-sequence** a sample and assemble the distinct genomes?
- I.e. "the whole metagenome"?
- This would enable a **very detailed** assessment of:
  - the organisms present
  - their phylotypic origin
  - the genes and functions present
  - I.e. provide in-depth **Community** and **Function** analysis
- The answer depends on the sequencing depth/coverage
  - So with enough sequence, in theory: "yes"
  - In practice, the answer is still "no" ***in general*** …
  - … but we will discuss this more later on

What is our data?

DNA sequence reads

So what do we sequence?

Development of sequencing methods to probe the Microbiome

- Early: sequencing of **whole “marker” genes** in clone libraries
  - Especially **16S ribosomal RNA genes** (SSU = small subunit)
  - These sequences can tell you **which species**, **genera etc** are present
  - Compared to today, this was very low throughput
  - Sequenced markers were long (~ **1,500 bp** 16S rRNA genes)
- Also direct sequencing of rRNA
- Sequencing of PCR products, e.g.
  - Boettger (1988) Rapid determination of bacterial ribosomal RNA sequences by direct sequencing of enzymatically amplified DNA. *FEMS Microbiol. Lett.* ***65:*** 171-176
  - Weisburg *et al.* (1991) 16s Ribosomal DNA Amplification for Phylogenetic Study *Journal of Bacteriology* **173** (2) 697-703

phylotypic barcode

5 901234 123457

Development of sequencing to probe the Microbiome

phylotypic barcode

- Modern sequencing methods are much faster/cheaper
  - But **reads are too short** to sequence **whole** 16S rRNA genes
  - At least, this is still the case for the sequencing platforms suitable for very low-error rate high-throughput
  - So, amplify and sequence the **most useful barcode region** of 16S rRNA genes
  - These variable regions of the gene identify the organisms present

- Another modern development – (“**shotgun**”) **metagenomics**
  - Sample and sequence genomic DNA at random
  - To as great a depth/coverage as feasible
  - These can identify the potential **functions** present (and organisms)
  - Can we assemble distinct, complete genomes from this data?

Sequencing phylotypic (“taxonomic”) marker genes

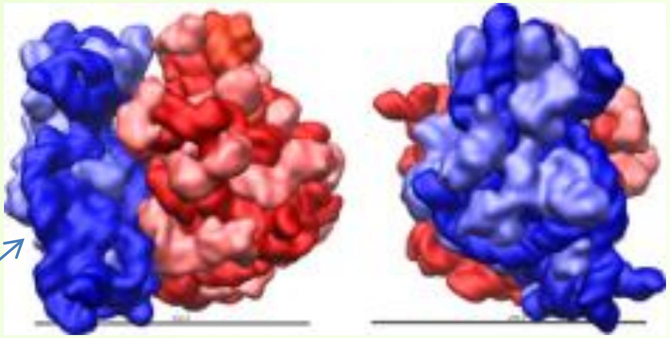
16S metagenomic amplicons (prokaryotes)
Ribosomal RNA barcodes
(Similar principles for 18S in eukaryotes;
28S/ITS regions often used in fungi)

Phylotypic barcode gene

- Must be present in all domains of life which you are investigating
- Must have extremely highly conserved regions to enable amplification
- Must have regions which mutate rapidly, to differentiate between organisms, and so:
  - **differ slightly between close relatives**
  - **differ a lot between distant relatives**

gene which codes for…

16S rRNA



Prokaryote ribosome

Red= LSU = large subunit (70S)

Blue = **SSU** = small subunit (30S)

Light blue = SSU protein

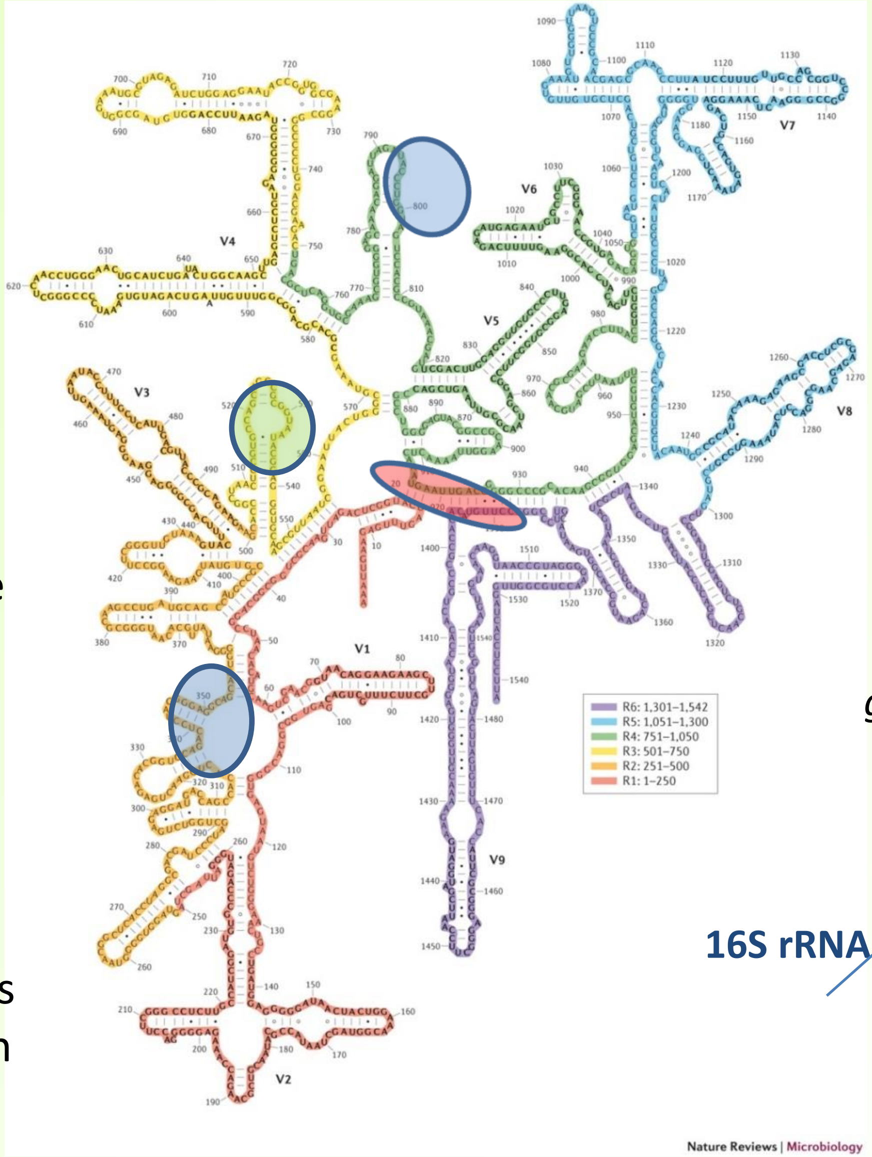
Dark blue = **SSU RNA = 16S rRNA**

Image:
Vossman,
Wikimedia
Commons

Amplification of a **segment** of the gene which codes for a **variable** region of the 16S rRNA molecule

→Primers

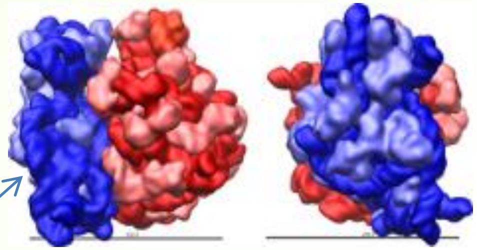
The variable region is chosen to distinguish between taxa



Nature Reviews | Microbiology

“barcode” for **taxa** (***phylotypes***)

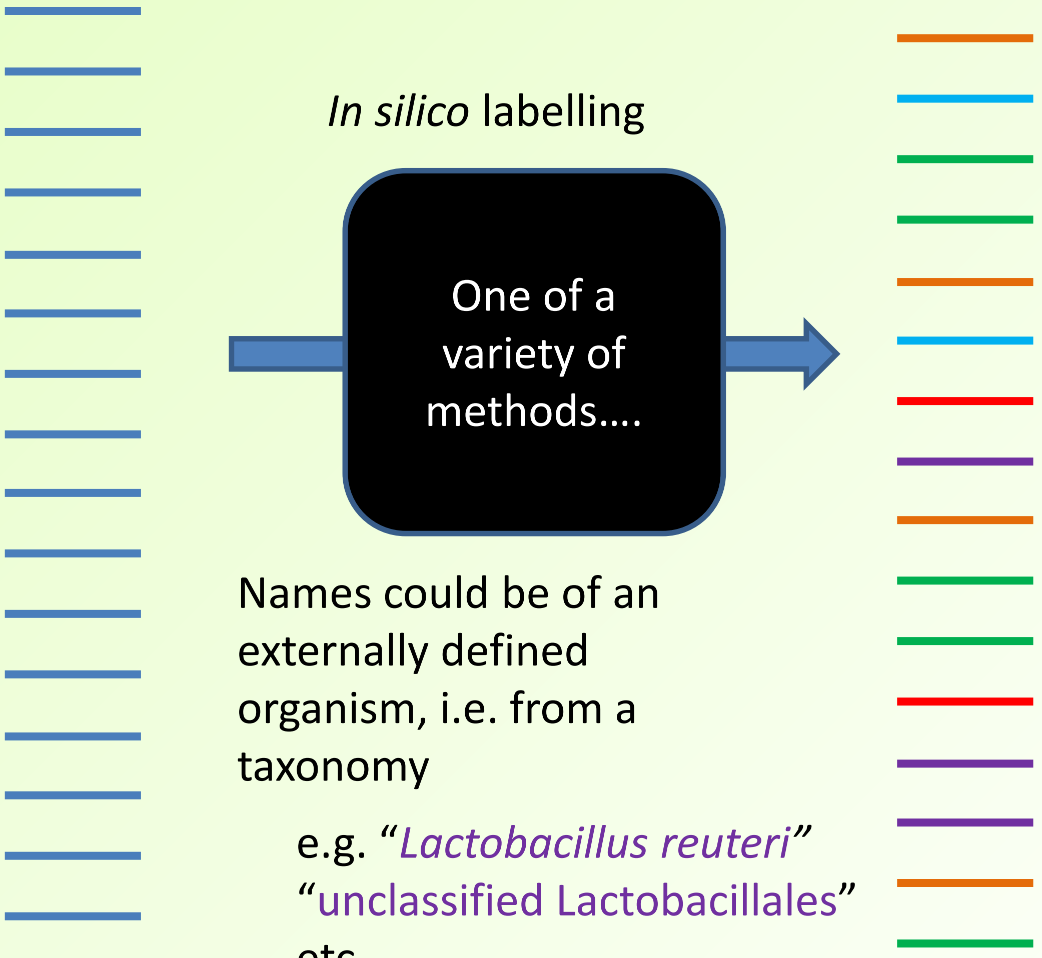
gene which codes for…



16S rRNA

Community analysis by marker-gene sequencing

Raw, unlabelled reads



In silico labelling

One of a variety of methods….

Label to indicate bug of origin

Name1

Name2

Name3

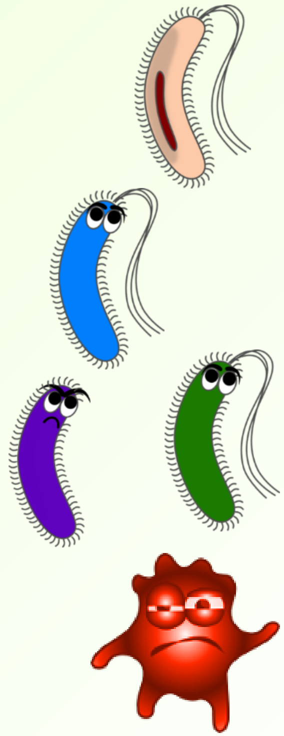
Name3

Name1

Name2

Name4

…etc..



Names could be of an externally defined organism, i.e. from a taxonomy

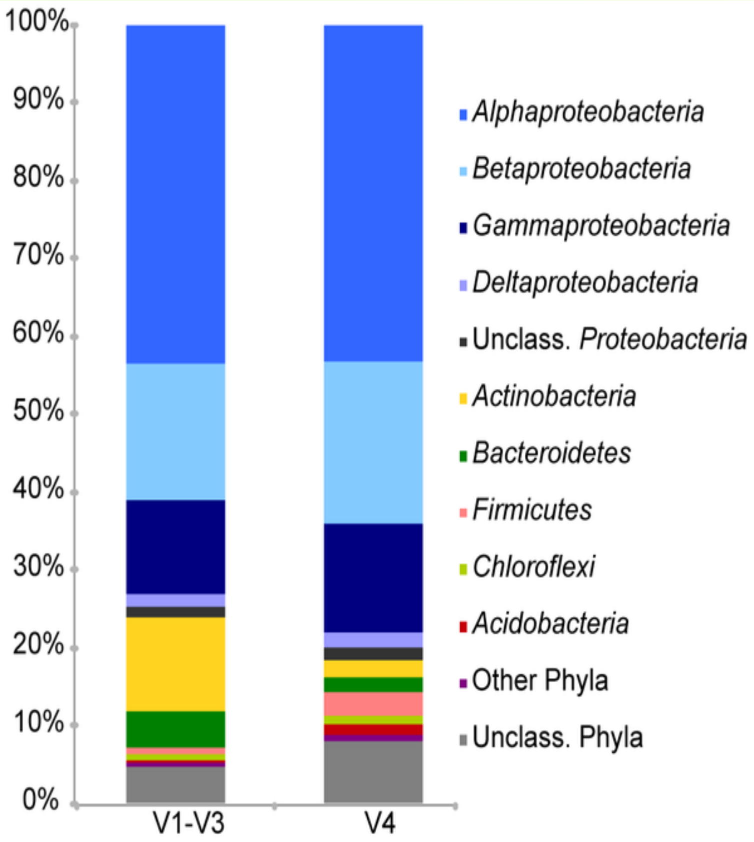
e.g. “*Lactobacillus reuteri*”
“unclassified Lactobacillales”
etc

Or could be **completely anonymous**, a name existing only within your data e.g. “OTU5432”
- Diversity studies still possible

Does it matter which 16S region you amplify?

- YES
- So, different amplified regions give you different results
- YES
- If you are interested in both Bacteria and Archaea, can you use the same primers for both?
- NO, not without introducing a lot of bias, in the general case
  - There are identified best available pairs for Archaea, and for Bacteria

Distribution of bacterial phyla and classes of Proteobacteria according to the 16S rRNA gene region.



Ibarbalz FM, Pérez MV, Figuerola ELM, Erijman L (2014) PLoS ONE 9(6): e99722. doi:10.1371/journal.pone.0099722

The Bias Associated with Amplicon Sequencing Does Not Affect the Quantitative Assessment of Bacterial Community Dynamics

Bias due to amplicon choice

- Is thought to be reproducible
- So you can compare like-with-like experiments
  - I.e. which amplified the same region
  - (strictly speaking, used the same primer pairs)
- There are other sources of bias/error
  - Mostly experimental stages, e.g.
    - Sample preparation
    - Sample storage
    - Sequencing platform itself (relatively low)
  - Potentially, Informatics – e.g. DB composition
  - More on this later…

This “*Who is in there?*” question…

What do you really want to know?

What are your questions?

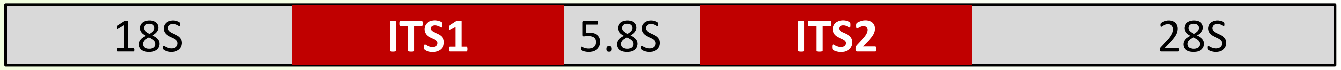
- How microbial **diversity** differs from one sample to another?
- Which **broad groups** (e.g. Phyla, Class) are present?
  - In what proportions?
  - How do these differ between samples?
- Narrower groups?
- Interest in particular **Species**, which may be generally abundant; or scarce across all samples?
- Hoping to find "smoking gun" microbes associated only with a particular condition?

Curated databases of rRNA gene sequences of taxonomic groups

- Catalogues of what the barcodes mean
  - **Ribosomal Database Project** (RDP) Cole *et al.* (2009) http://rdp.cme.msu.edu/
    - **16S** bacteria+archaea (small subunit)
    - **28S** fungi (large subunit)
  - **Greengenes** DeSantis *et al.* (2006) http://greengenes.lbl.gov/
    - **16S** bacteria+archaea (small subunit)
  - **SILVA** Pruesse *et al.* (2007) http://www.arb-silva.de/
    - **16S** bacteria+archaea (small subunit)
    - **18S** eukaryote (small subunit)
    - **23S** bacteria+archaea (large subunit)
    - **28S** eukaryote (large subunit)

Marker genes: variants (1)

- **Eukaryote ribosomal genes:**
- Fungal **18S rRNA gene**
  - E.g. Lumini *et al.* (2009) Disclosing arbuscular mycorrhizal fungal biodiversity in soil through a land-use gradient using a pyrosequencing approach *Environmental Microbiology **12*** (8) 2165-79
  - Choice of 18S was due to the (then) limits of the read lengths, rendering more traditional ITS sequencing less useful
- Internal Transcribed Spacers of nuclear rRNA gene (ITS1, ITS2)
  - the default approach for fungi



DBs of fungal ITS/rRNA sequences

- UNITE : ITS sequences https://unite.ut.ee/
  - Kõljalg *et al.* (2013) Towards a unified paradigm for sequence-based identification of fungi *Molecular Ecology* **22** (21) 5271-7
- SILVA has eukaryote SSU, LSU sequences
- RDP has fungal 28S sequences

Marker genes: variants (2)

- Other taxonomic marker genes, e.g:
- *amoA* gene encodes an ammonia monooxygenase subunit
  - Present in bacteria and archaea
  - Leininger *et al.* (2006) Archaea predominate among ammonia-oxidizing prokaryotes in soils *Nature* **442** 806-9
  - this study used different primers for bacteria and archaea
  - in tandem with soil “meta-lipidomics”
- Increasing use of other protein-coding marker genes, e.g.
  - rpoB (RNA polymerase beta subunit)
  - Others e.g. rplB, pyrG, fusA, leuS
  - Some studies have indicated they are consistent with 16S results but may provide deeper resolution

Metatranscriptomics and 16S sequences

- If you sample/sequence “all” the metatranscriptome
  - you get mostly rRNA
  - cells make loads of ribosomes!
- **This can be used for community analysis**
- The (relatively small) amount of mRNA can be used simultaneously for functional studies
- In practice, a metatranscriptomics study is likely to target a particular aspect such as expression of protein-coding genes
  - So would be enriched for mRNA

Metagenomics and marker genes

- Traditionally, all metagenome reads would be compared with a reference database
  - By some method or another
  - Taxonomic (and functional) labels
- Tools now exist which enable more targeted evaluation of the most useful marker genes
  - Which will be/may be present in your data
  - More on this later

Marker genes: the functional twist

- If your marker-gene sequencing identifies characterised organisms with known functions
- Then can we use quantitative marker results to infer microbiome functions, quantitatively?
- A very qualified “yes”
  - – PICRUSt: Langille *et al.* (2013) Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences *Nature Biotechnology* **31** 814-21
    - “with quantifiable uncertainty”

Picture credits

- (space shuttle) By NASA https://commons.wikimedia.org/w/index.php?curid=137540
- By Office of Surface Mining - http://www.osmre.gov/sovern.htm, Public Domain, https://commons.wikimedia.org/w/index.php?curid=4267418
- (hands) By Ibex73 - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=47937549
- (GIT) By Edelhart Kempeneers - Gray's Anatomy, Public Domain, https://commons.wikimedia.org/w/index.php?curid=534843
- (nasal/oral anatomy) By Internet Archive Book Images -
- https://commons.wikimedia.org/w/index.php?curid=43370262
- (termite) By CSIRO, CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=35479115