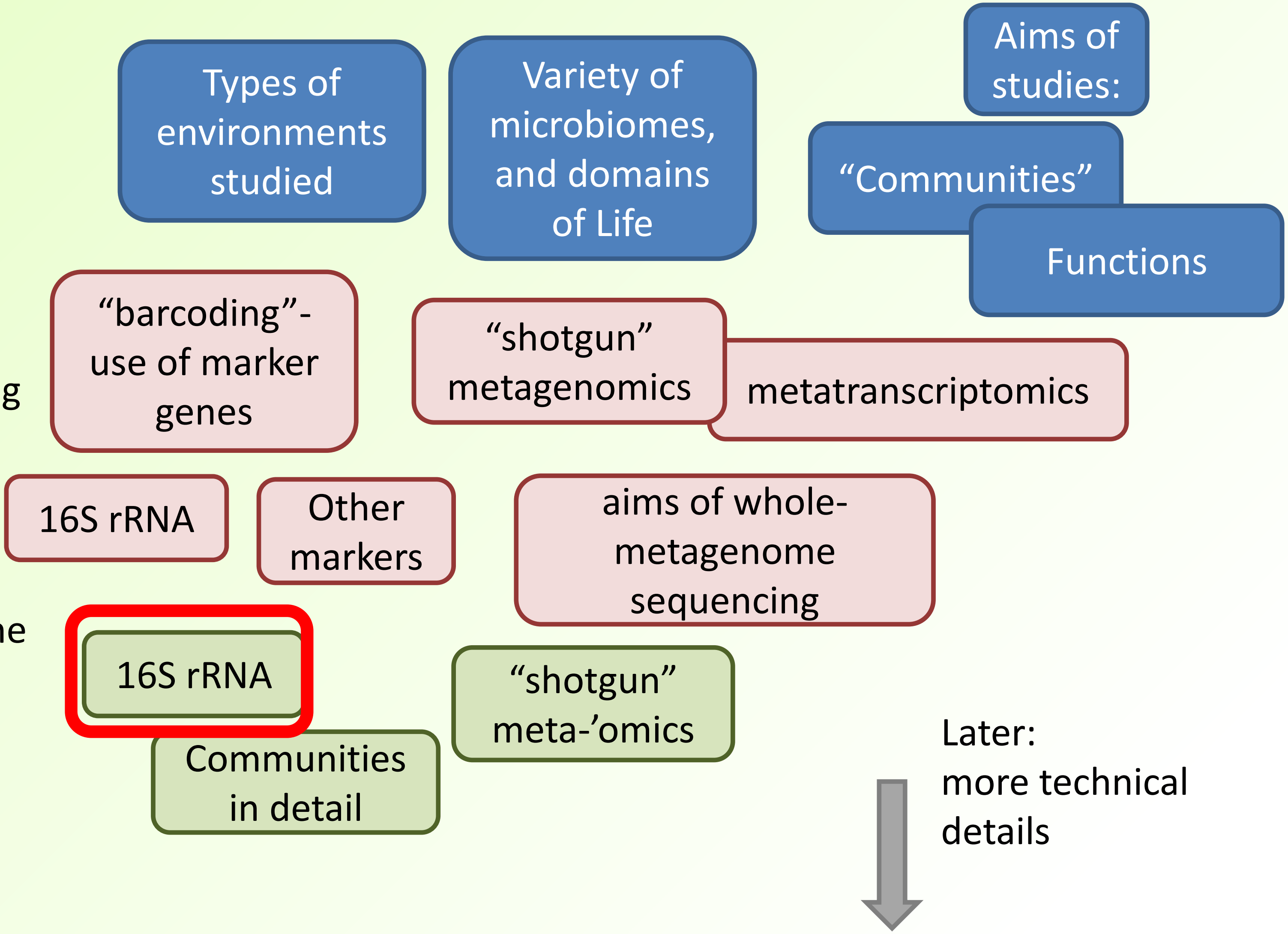
Introducing Microbiome Bioinformatics

Part 5.

Recap: Aims

- **Microbiome analysis**
  - with particular regard to **sequence informatics concepts**
- "Top down" – putting analysis tools and resources in context
- No highly detailed technicalities (yet)
  - No instructions on how to run particular programs
- Why you are using the bioinformatics approaches you use; pros, cons; alternatives

Topics, top-down



Introduction

Types of environments studied

Variety of microbiomes, and domains of Life

Aims of studies:

"Communities"

Functions

Aims of DNA/ RNA sequencing approaches

"barcoding"- use of marker genes

"shotgun" metagenomics

metatranscriptomics

16S rRNA

Other markers

aims of whole-metagenome sequencing

Dealing with the data: Informatics concepts

16S rRNA

Communities in detail

"shotgun" meta-'omics

Later: more technical details

Series of talks

- 4 so far
- Open ended… as long there is demand
- Expected to be every 2 weeks, but all dates will be confirmed in advance
  - *Bite-size bioinformatics mailing list*
- The next few will cover: (*not necessarily in this order…)*
  - 16S analysis for community profiling
  - Classification issues (taxonomies etc)
  - Analysing richness and diversity of those communities
  - Dealing with sequencing and other errors
- Informal and flexible
  - Please interrupt and ask questions
  - Suggestions for topics for further focus

Series of talks

- Part 1: 27/1/2017
  - "Biological and Experimental Stuff that a microbiome bioinformatician needs to know"
  - Overview of marker gene sequencing for community analysis
- Part 2: 10/2/2017
  - Overview of whole-metagenome sequencing
- Part 3: 24/2/2017
  - Focus on metatranscriptomics
- Part 4: 10/3/2017
  - Different bioinformatics approaches to processing 16S read data
- Slideshows
  - http://ghfs1.ifr.ac.uk/ghfs/

To be confirmed…

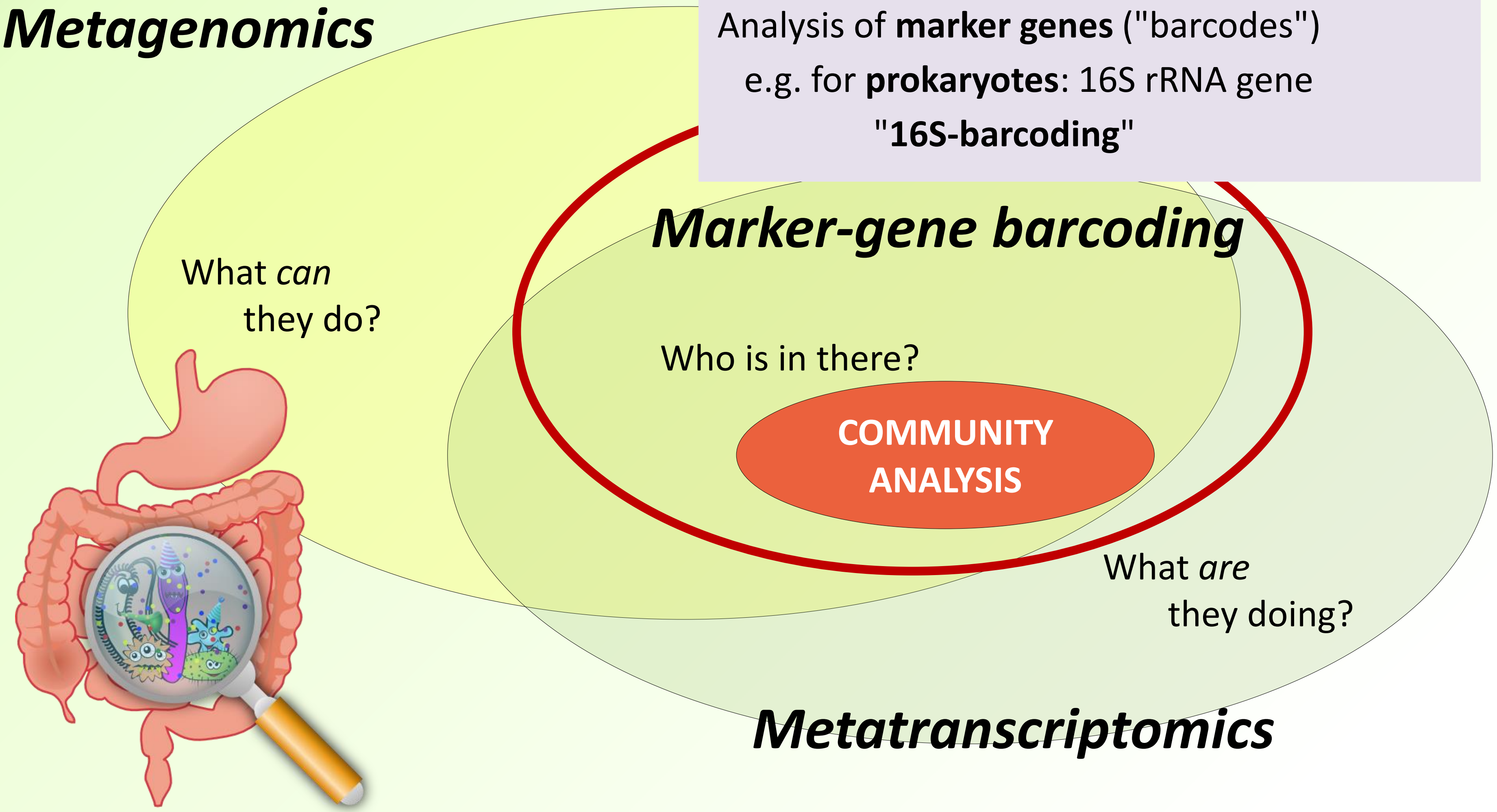
7th April	Barton
21st April	Rollesby
5th May	Barton
~~19th May~~	~~Rollesby~~ cancelled

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

Analysis of **marker genes** ("barcodes")

e.g. for **prokaryotes**: 16S rRNA gene

"**16S-barcoding**"



Metagenomics

What *can* they do?

Marker-gene barcoding

Who is in there?

COMMUNITY ANALYSIS

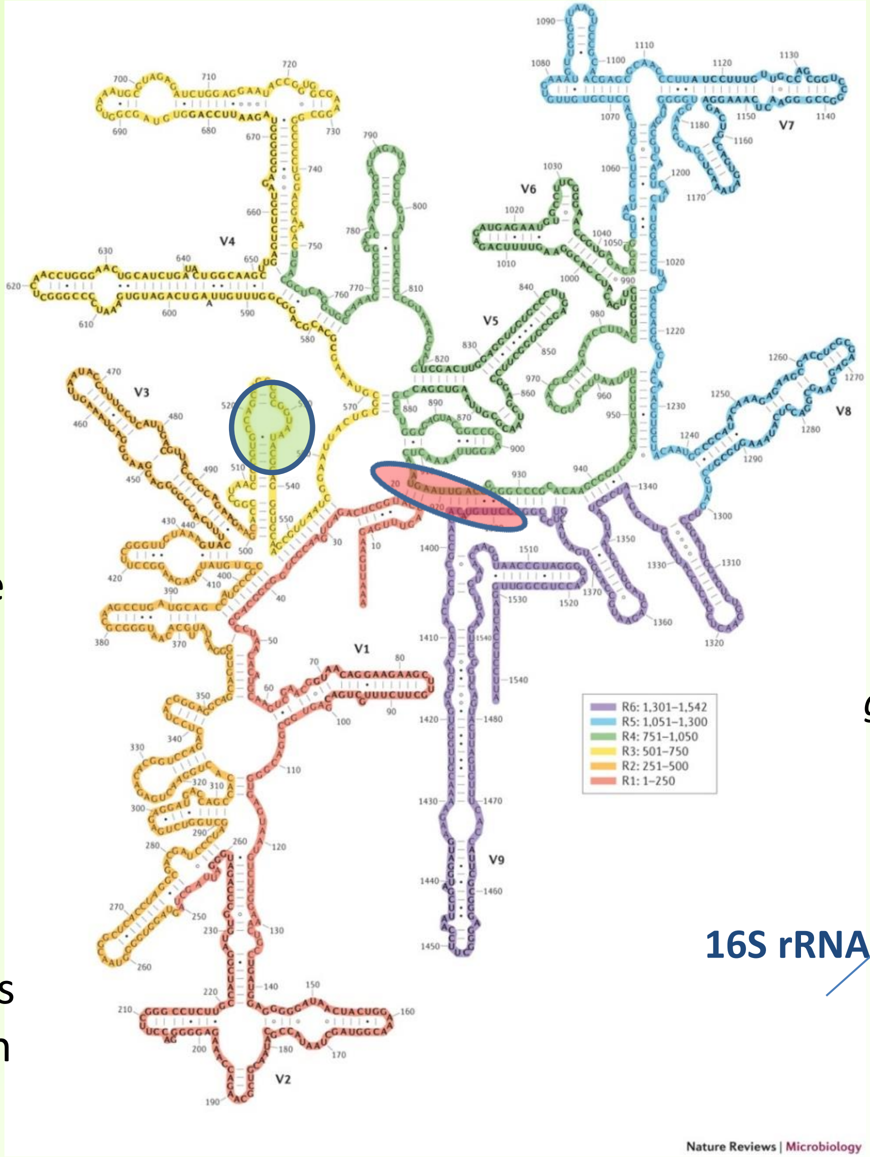
What *are* they doing?

Metatranscriptomics

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

→Primers

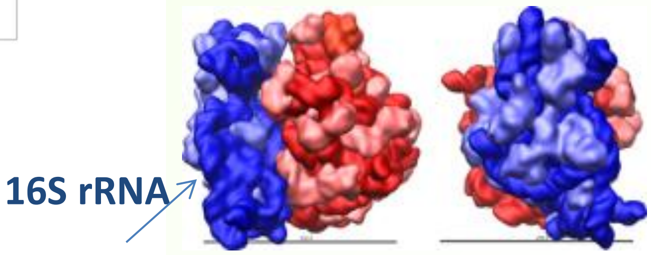
The variable region is chosen to distinguish between taxa



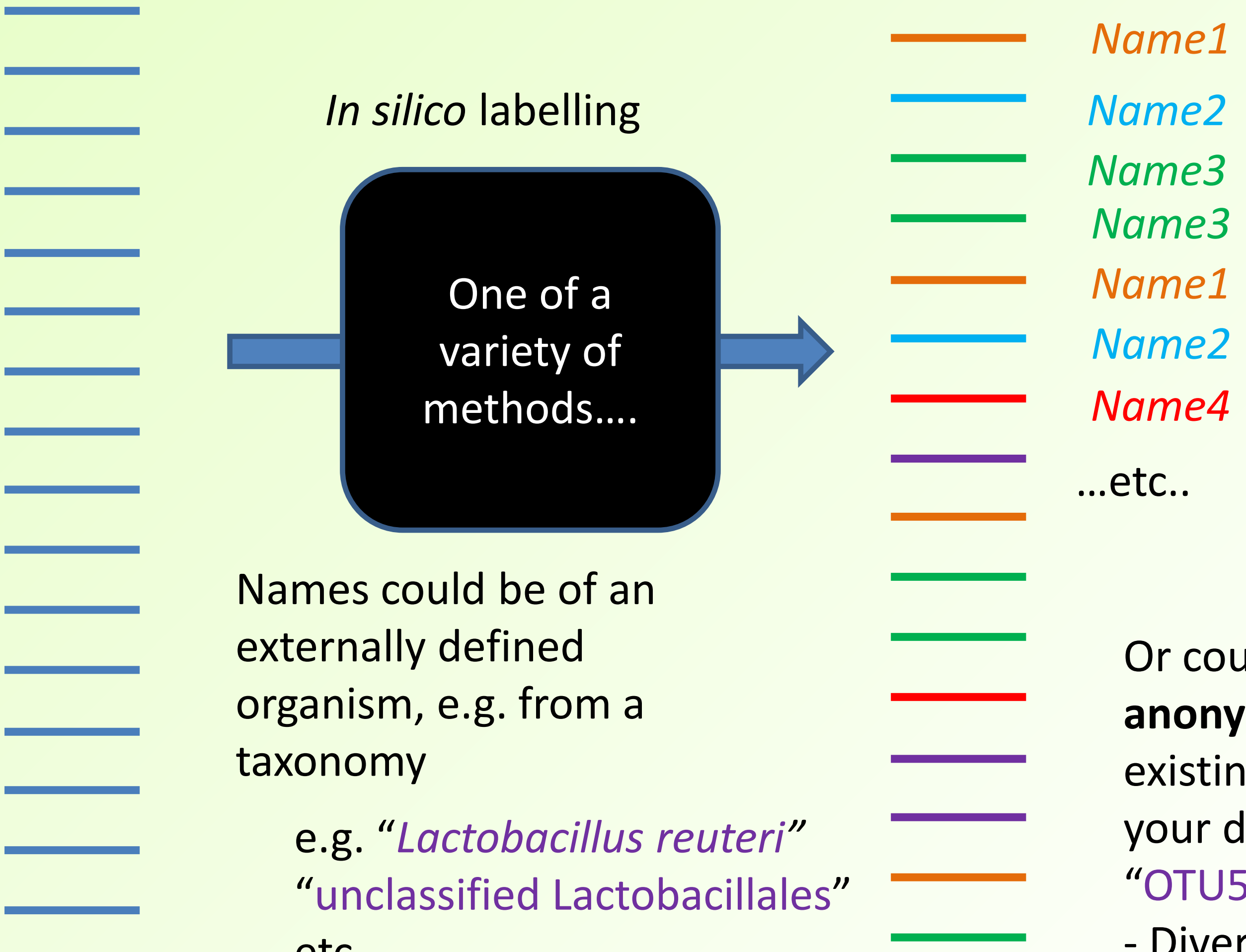
Nature Reviews | Microbiology

marker gene
("barcode")
for ***phylotypes***

gene which codes for…



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

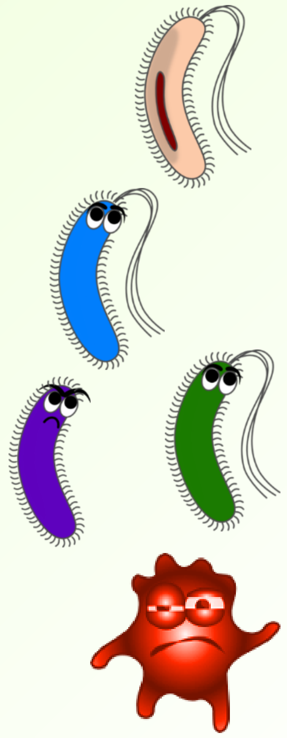
Name3

Name1

Name2

Name4

…etc..



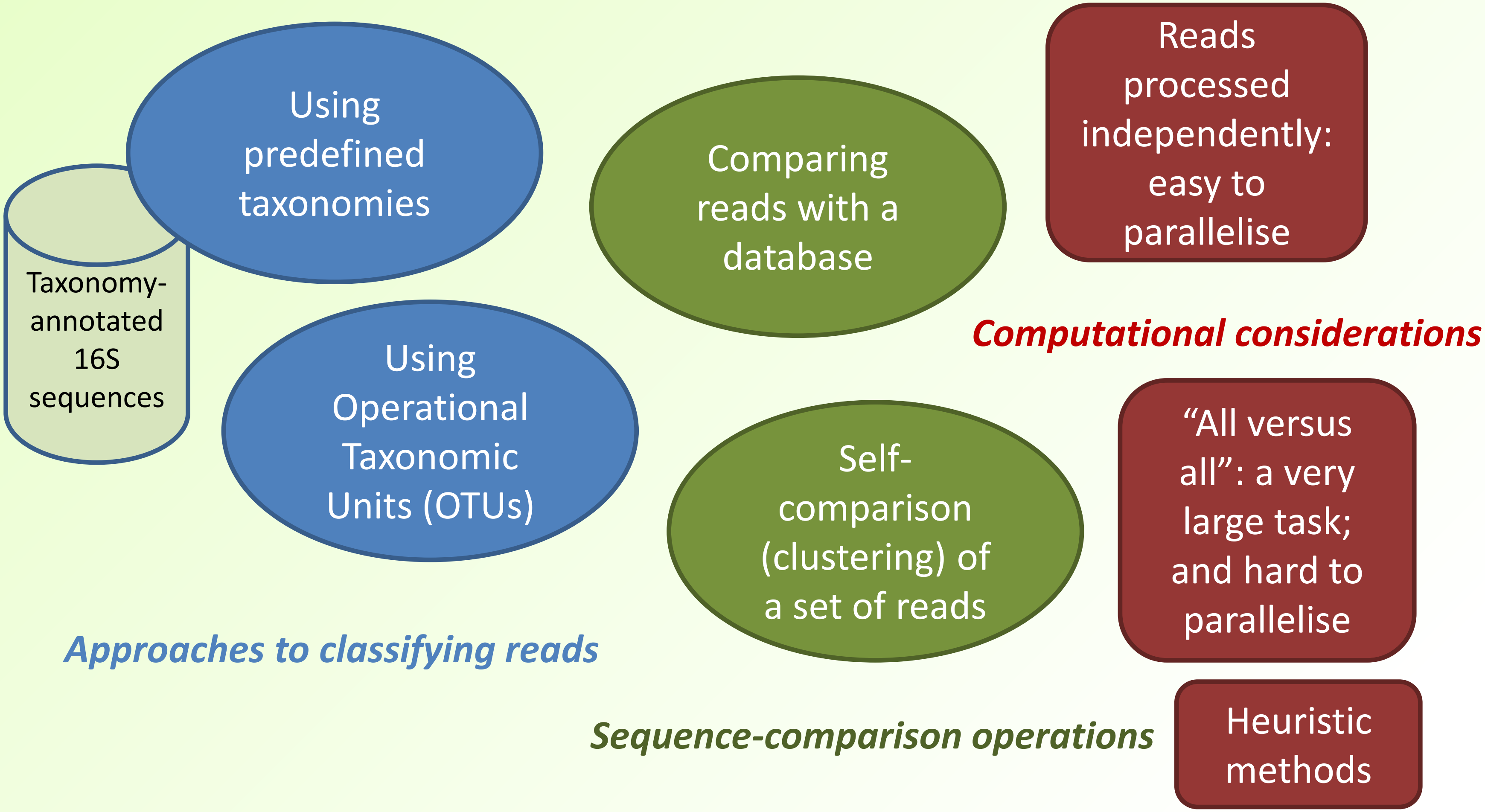
Names could be of an externally defined organism, e.g. from a taxonomy

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

Recap- some considerations

(not mutually exclusive)



Taxonomy-annotated 16S sequences

Using predefined taxonomies

Using Operational Taxonomic Units (OTUs)

Approaches to classifying reads

Comparing reads with a database

Self-comparison (clustering) of a set of reads

Sequence-comparison operations

Reads processed independently: easy to parallelise

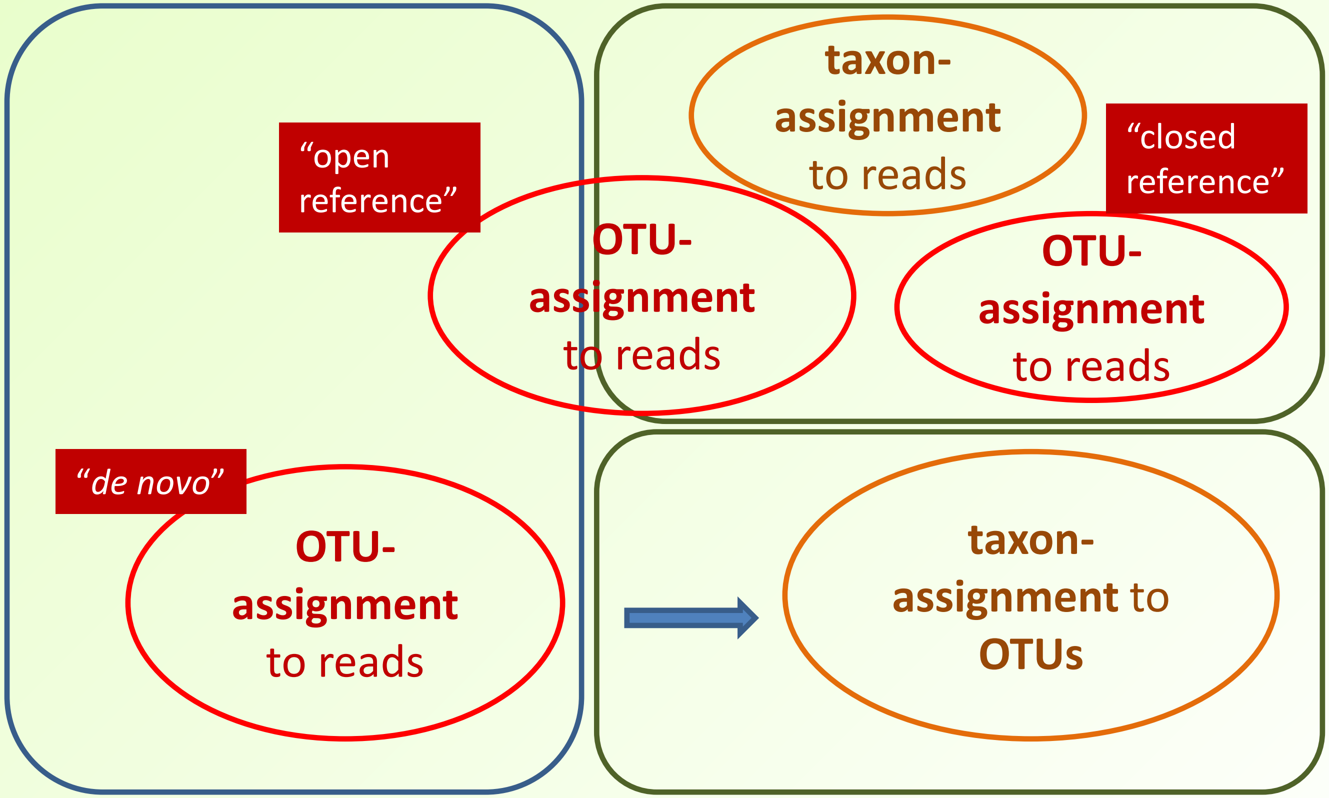
Computational considerations

“All versus all”: a very large task; and hard to parallelise

Heuristic methods

Clustering : comparing reads with each other ("self-referential")

Using a reference database



"open reference"

OTU-assignment to reads

taxon-assignment to reads

"closed reference"

OTU-assignment to reads

"*de novo*"

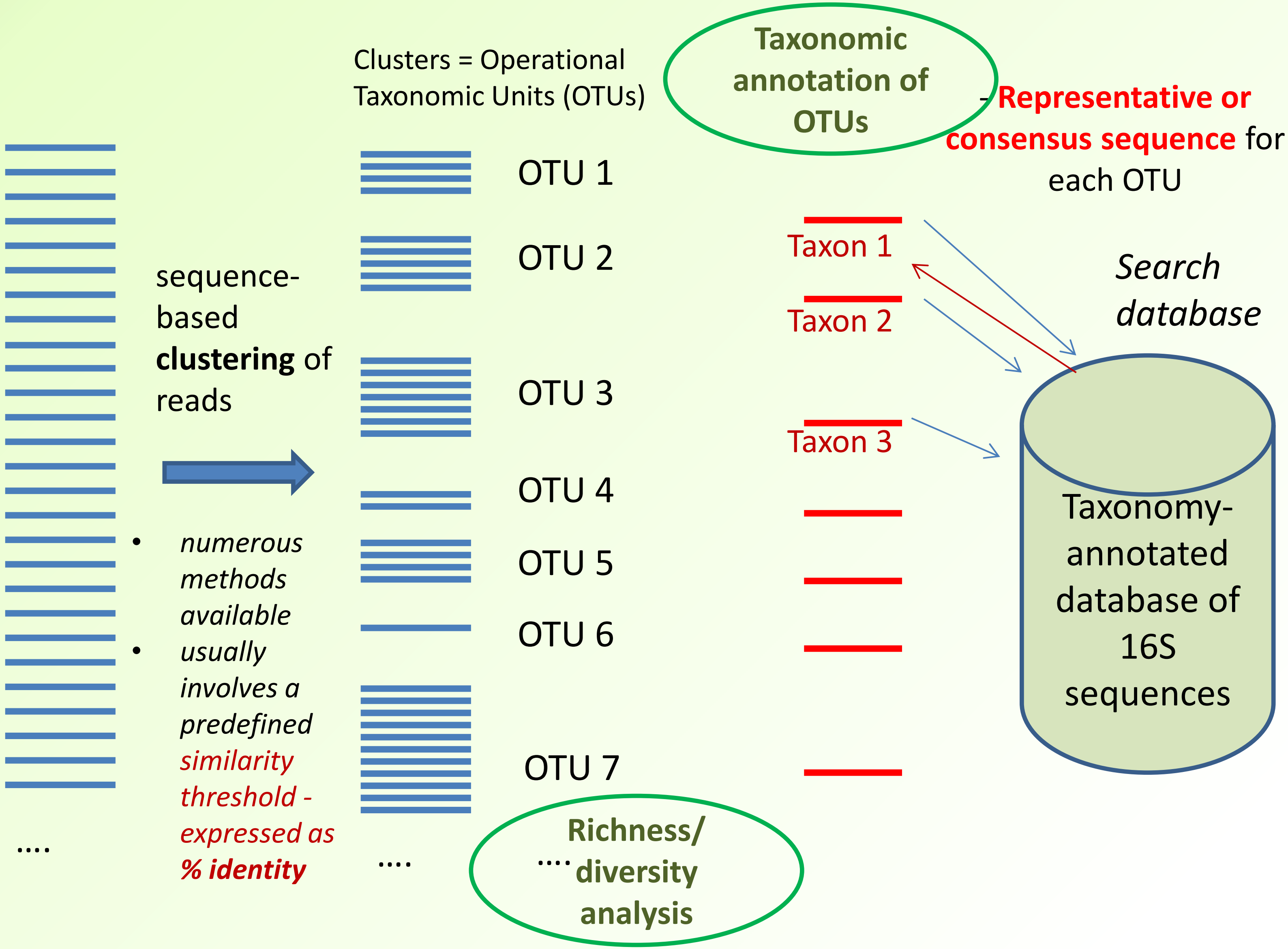
OTU-assignment to reads

taxon-assignment to **OTUs**

processing each **read** independently

processing representative sequences of each OTU

OTUs by *de novo* clustering (not the only way)



Clusters = Operational Taxonomic Units (OTUs)

Taxonomic annotation of OTUs

- **Representative or consensus sequence** for each OTU

sequence-based **clustering** of reads

- *numerous methods available*
- *usually involves a predefined similarity threshold - expressed as **% identity***

OTU 1

OTU 2

OTU 3

OTU 4

OTU 5

OTU 6

OTU 7

....

....

....

Taxon 1

Taxon 2

Taxon 3

Search database

Taxonomy-annotated database of 16S sequences

Richness/ diversity analysis

Sequence Identity

16S rRNA gene: V4-V5 region

94.5% sequence identity in **V4-V5** region

Escherichia coli K-12 MC1400
Enterobacter cloacae

(same Family)

```
E._coli_K-12         1 TACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCA     50
                       ||||||||||||||||||||||||||||||||||||||||||||||||||
Enterobacter         1 TACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCA     50

E._coli_K-12        51 GGCGGTTTGTTAAGTCAGATGTGAAATCCCCGGGCTCAACCTGGGAACTG    100
                       ||||||.|||.|||||.|||||||||||||||||||||||||||||||||
Enterobacter        51 GGCGGTCTGTCAAGTCGGATGTGAAATCCCCGGGCTCAACCTGGGAACTG    100

E._coli_K-12       101 CATCTGATACTGGCAAGCTTGAGTCTCGTAGAGGGGGTAGAATTCCAGG    150
                       |||..||.|||||||.|||.||||||.|||||||||||||||||||||||
Enterobacter       101 CATTCGAAACTGGCAGGCTAGAGTCTTGTAGAGGGGGTAGAATTCCAGG    150

E._coli_K-12       151 TGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGCGAAGGCG    200
                       ||||||||||||||||||||||||||||||||||||||||||||||||||
Enterobacter       151 TGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGCGAAGGCG    200

E._coli_K-12       201 GCCCCCTGGACGAAGACTGACGCTCAGGTGCGAAAGCGTGGGGAGCAAAC    250
                       |||||.|||||.|||||||||..||||||||.||||||||||||||||||
Enterobacter       201 GCCCCTTGGACAAAGACTGACCTTCAGGTGCCAAAGCGTGGGGAGCAAAC    250

E._coli_K-12       251 AGG    253
                       |||
Enterobacter       251 AGG    253
```

(same Phylum, different Classes)

75.1% sequence identity in V4-V5 region *

Escherichia coli K-12 MC1400
Campylobacter jejuni SSI 5384-98

```
E._coli_K-12       1 TACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCA     50
                     ||||||||||||||||||||.|||||||.|||||||||||||.||.||.|
Campylobacter      1 TACGGAGGGTGCAAGCGTTACTCGGAATCACTGGGCGTAAAGGGCGCGTA     50

E._coli_K-12      51 GGCGGTTTGTTAAGTCAGATGTGAAATCCCCGGGCTCAACCTGGGAACTG    100
                     |||||.||.|.|||||...|||||||||....||||.||||....|||||
Campylobacter     51 GGCGGATTATCAAGTCTCTTGTGAAATCTAATGGCTTAACCATTAAACTG    100

E._coli_K-12     101 CATCTGATACTGGCAAGCTTGAGTCTCGTAGAGGGGGGTAGAATTCCAGG    150
                     |.|..||.||||..|..||.||||...|.|||||..|.|.|||||...||
Campylobacter    101 CTTGAGAAACTGATAGTCTAGAGTGAGGGAGAGGCAGATGGAATTGGTGG    150

E._coli_K-12     151 TGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGCGAAGGCG    200
                     |||||.|||.||||.||||||.|||...|.|||||||..|.|||||||||
Campylobacter    151 TGTAGGGGTAAAATCCGTAGATATCACCAAGAATACCCATTGCGAAGGCG    200

E._coli_K-12     201 GCCCCCTGGACGAAGACTGACGCTCAGGTGCGAAAGCGTGGGGAGCAAAC    250
                     .....|||||.....|||||||||.|||.|||||||||||||||||||||
Campylobacter    201 ATTTGCTGGAACTCAACTGACGCTAAGGCGCGAAAGCGTGGGGAGCAAAC    250

E._coli_K-12     251 AGG    253
                     |||
Campylobacter    251 AGG    253
```

* With this particular scoring scheme
% SEQUENCE IDENTITY IS NOT AN IMMUTABLE PROPERTY OF A PAIR OF SEQUENCES

(different Phyla)

79.1% sequence identity in V4-V5 region

Escherichia coli K-12 MC1400
Lactobacillus salivarius JCM 1231

```
E._coli_K-12        1 TACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCA     50
                      ||||.|||..||||||||||..||||.|||.|||||||||||.|.|||||
Lactobacillus       1 TACGTAGGTGGCAAGCGTTATCCGGATTTATTGGGCGTAAAGGGAACGCA     50

E._coli_K-12       51 GGCGGTTTGTTAAGTCAGATGTGAAATCCCCGGGCTCAACCTGGGAACTG    100
                      ||||||.|.|||||||.|||||||||.||...||||.||||.|.|.|.||
Lactobacillus      51 GGCGGTCTTTTAAGTCTGATGTGAAAGCCTTCGGCTTAACCGGAGTAGTG    100

E._coli_K-12      101 CATCTGATACTGGCAAGCTTGAGTCTCGTAGAGGGGGGTAGAATTCCAGG    150
                      |||..||.|||||.|..|||||||...|.|||||.|.||.|||.||||.|
Lactobacillus     101 CATTGGAAACTGGGAGACTTGAGTGCAGAAGAGGAGAGTGGAACTCCATG    150

E._coli_K-12      151 TGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGCGAAGGCG    200
                      |||||||||||||||||||||.||.||||.|||.|||.||||||||.|||
Lactobacillus     151 TGTAGCGGTGAAATGCGTAGATATATGGAAGAACACCAGTGGCGAAAGCG    200

E._coli_K-12      201 GCCCCCTGGACGAAGACTGACGCTCAGGTGCGAAAGCGTGGGGAGCAAAC    250
                      ||.|.||||.|....|||||||||.||||.||||||||||||.|||||||
Lactobacillus     201 GCTCTCTGGTCTGTAACTGACGCTGAGGTTCGAAAGCGTGGGTAGCAAAC    250

E._coli_K-12      251 AGG    253
                      |||
Lactobacillus     251 AGG    253
```

(different Domains/Kingdoms)

64.3% sequence identity in V4-V5 region

Escherichia coli K-12 MC1400
Methanobrevibacter acididurans ATM

```
E._coli_K-12       1 -TACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCACGC     49
                      ..|||..|.| |.||.|.||...|...|||.|||||.|||||||..||.
Methanobrevib      1 ACCCGGCAGCT-CTAGTGGTAGCTGTTTTTATTGGGCCTAAAGCGTTCGT     49

E._coli_K-12      50 AGGCGGTTTGTTAAGTCAGATGTGAAATCCCCGGGCTCAAC-CTGGGAAC     98
                     ||.||||||..||||||....|||||||||......|.||| .||||||.
Methanobrevib     50 AGCCGGTTTAATAAGTCTTTGGTGAAATCCTGTTTTTTAACTATGGGAAT     99

E._coli_K-12      99 TGCATCTGATACTGGCAAGCTTGAGTCTCGTAGAGGGGGTAGAATTCCA    148
                     |||....|||||||..|.|||||||....|.|||||...|..|.|.|||.
Methanobrevib    100 TGCTGAGGATACTGTTAGGCTTGAGGTCGGGAGAGGTTAGCGGTACTCCC    149

E._coli_K-12     149 GGTGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGCGAAGG    198
                     .|.||||.||||||||.|....|.....||||||..|||.||||||||||
Methanobrevib    150 AGGGTAGGGGTGAAATCCTGTAATCCTGGGAGGACCACCTGTGGCGAAGG    199

E._coli_K-12     199 CGGCCCCCTGGACGAAGACTGACGCTCAGGTGCGAAAGCGTGGGGAGCAA    248
                     ||||...|||||...|..||||||.|.|||..|||||||..||||.||.|
Methanobrevib    200 CGGCTAACTGGAACGAACCTGACGGTGAGGGACGAAAGCTAGGGGCGCGA    249

E._coli_K-12     249 ACAGG    253
                     ||.||
Methanobrevib    250 ACCGG    254
```

Clustering algorithms

- A complex topic - more on this later on
- For now, just be aware that sequence-clustering algorithms for OTU-assignment usually use a similarity threshold
  - **Expressed as a percentage sequence identity** – usually 97%
- For any given threshold, the results depend on which clustering algorithm is used
- % identity thresholds are also highly relevant in other sequence-comparison contexts (besides clustering) for dealing with OTUs

97 ….why?

- "Almost all published papers use 97% clustering, so this will be easier to explain to your PI and to referees."
  - – Robert Edgar, UPARSE/UNOISE FAQ, drive5.com

- That's good for consistency
- Although the algorithm-dependent results for any *x*%, is not

- …. but where did 97% come from?

Guidelines

- Using 97% is fine
- Don’t expect your OTUs to equate to “species”
  - Or any other predefined taxonomic level
- Using 97% gives perhaps the best chance of comparability with other published studies
- in any case But: (ir)reproducibility of results is very depend on other things– such as:
  - Which clustering/assignment algorithm is used
  - How amplification/sequencing errors are handled

So, what ***are*** OTUs?

What do they represent?

How do they relate to taxa?

…and what's so special about the number 97?

A very brief summary of taxonomy

More details of prokaryote taxonomy – and why it’s sometimes quite annoying - in a future session

Strictly, “Systematics”

- The discipline of taxonomic classification

Taxonomies

Some oft-used taxonomic levels	**Example taxa**
Kingdom (Domain)	Bacteria
Phylum	Firmicutes
Class	Bacilli
Order	Lactobacillales
Family	Streptococcaceae
Genus	*Streptococcus*
Species	*Streptococcus pneumoniae*
strain	*S. pneumoniae* ATCC 700669

“higher” (more inclusive)

(Numerous intermediate levels are also used – not shown)

“lower” (more exclusive)

- Organisms classified in the same taxa share:
  - Characteristics (observable)
  - Common descent (inferred)
- Organisms in the lowest groups share the most characteristics
  - And are the most recently diverged
  - **Species** represent **isolated reproductive groups**
- That's the idea anyway…
- But taxonomy is difficult…. and messy

“there are as many ideas on species as there are biologists”

- Cowan (1968) *A dictionary of microbial taxonomic usage*, pub. Oliver and Boyd
- ‘Mayden (1997) categorized the **25 concepts** developed until 1996 and arranged them in a hierarchical order’
  - Hohenegger (2012) Transferability of genomes to the next generation: the fundamental criterion of the biological species, *Zootaxa* **3572** 11-17
  - (Ref: Mayden, R.L.: A Hierarchy of species concepts: the denouement in the saga of the species problem, *In: Species: The Units of Biodiversity*, Claridge, Dawah, Wilson (Eds.) Chapman and Hall, 1997)

evolutionary

recognition

nominalistic

phylogenetic

biological

monophyletic

phenetic

…and many more…

Some are ‘operational’

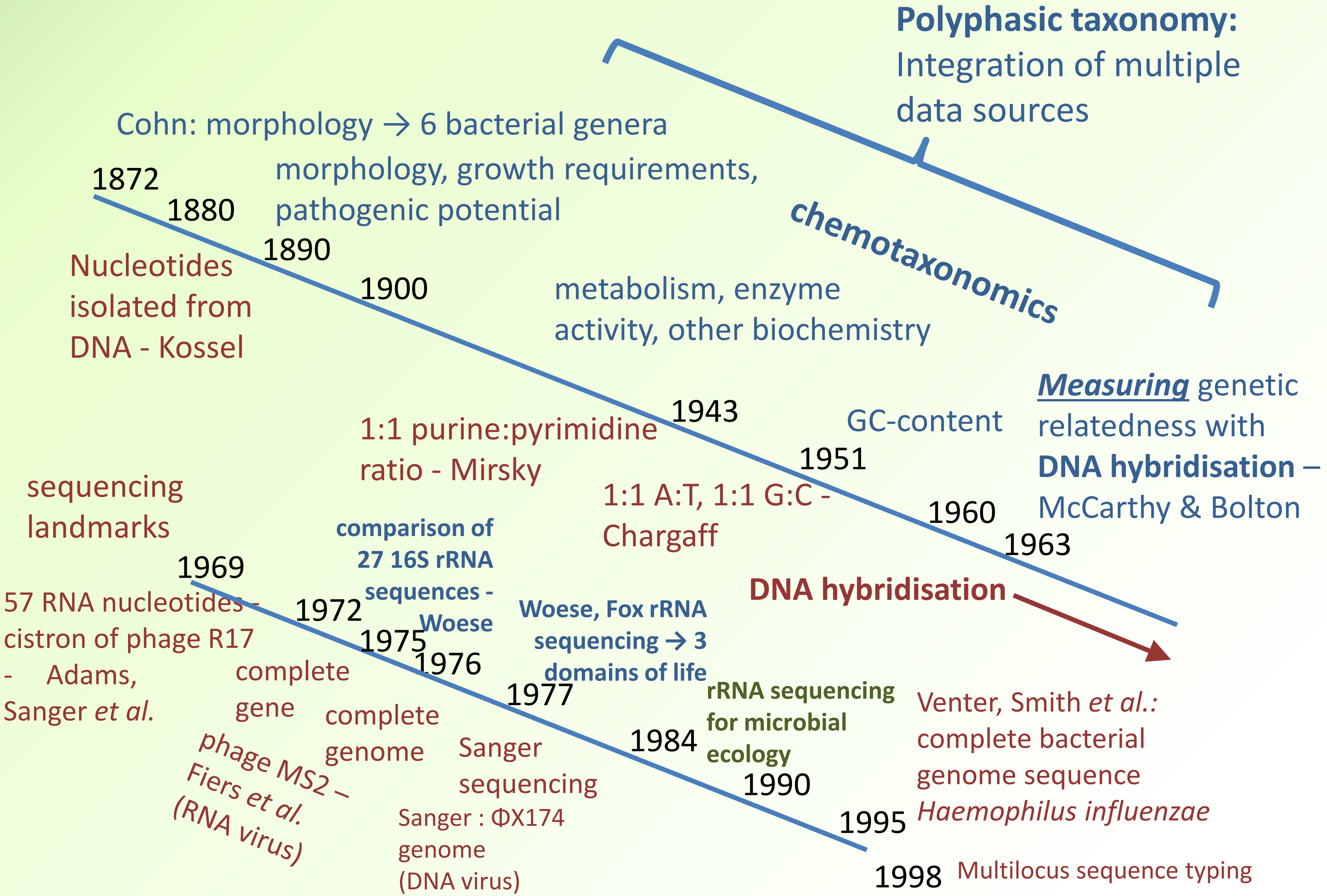
Some are not

Some are ‘explanatory’

Some are not

How did we get here?

Even just with prokaryotes, it’s complicated enough



Polyphasic taxonomy: Integration of multiple data sources

Cohn: morphology → 6 bacterial genera

1872

1880

morphology, growth requirements, pathogenic potential

1890

Nucleotides isolated from DNA - Kossel

1900

metabolism, enzyme activity, other biochemistry

chemotaxonomics

1943

1:1 purine:pyrimidine ratio - Mirsky

GC-content

1951

1:1 A:T, 1:1 G:C - Chargaff

1960

1963

Measuring genetic relatedness with **DNA hybridisation** – McCarthy & Bolton

DNA hybridisation

sequencing landmarks

1969

57 RNA nucleotides - cistron of phage R17 - Adams, Sanger *et al.*

1972

complete gene

phage MS2 – Fiers *et al.* (RNA virus)

comparison of 27 16S rRNA sequences - Woese

1975

complete genome

1976

Sanger : ΦX174 genome (DNA virus)

Woese, Fox rRNA sequencing → 3 domains of life

1977

Sanger sequencing

1984

rRNA sequencing for microbial ecology

1990

1995

Venter, Smith *et al.*: complete bacterial genome sequence *Haemophilus influenzae*

1998 Multilocus sequence typing

DNA-Hybridisation, sequence identity and taxonomy

Brought to you by the numbers **70**, **5** and **97**

Backward compatibility

- DNA-DNA hybridisation was found to be **consistent** with results using **established taxonomic criteria**
- Providing greater **resolution**
- Enabled (for closely-related organisms) a **measurement** of the amount of DNA which hybridises (relative to self-hybridisation)

DNA-DNA hybridisation: units

- For any pair of organisms, the hybridisation is expressed as a single number:
- Basically, a **proportion** (expressed as a **percentage**) of the amount of DNA which binds
  - ***relative** to self-hybridisation* under the same conditions.
- Let's refer to this as **"% relative binding**"
  - (known by many other names, confusingly)

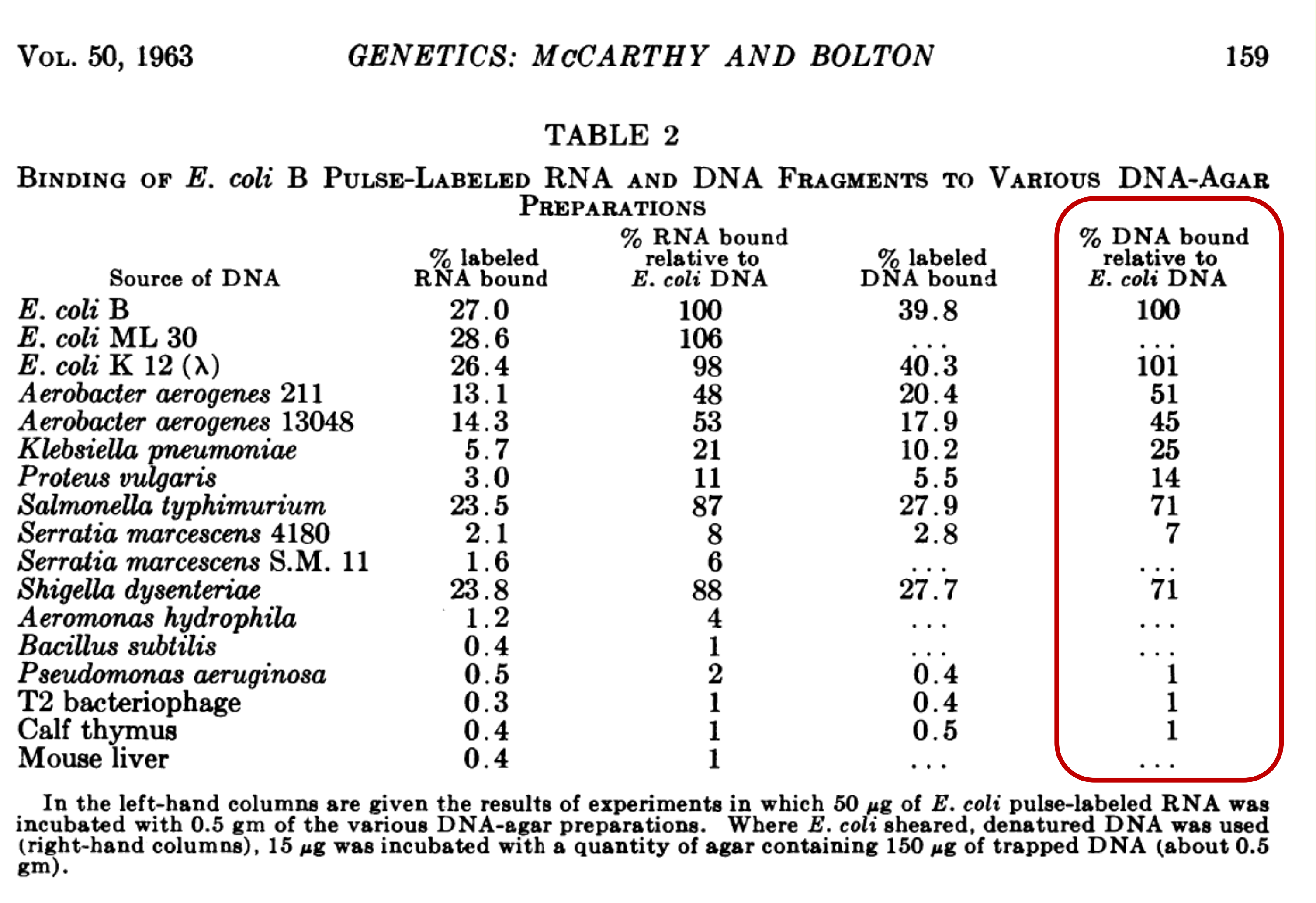


TABLE 2

BINDING OF *E. coli* B PULSE-LABELED RNA AND DNA FRAGMENTS TO VARIOUS DNA-AGAR PREPARATIONS

Source of DNA	% labeled RNA bound	% RNA bound relative to *E. coli* DNA	% labeled DNA bound	% DNA bound relative to *E. coli* DNA
E. coli B	27.0	100	39.8	100
E. coli ML 30	28.6	106	...	...
E. coli K 12 (λ)	26.4	98	40.3	101
Aerobacter aerogenes 211	13.1	48	20.4	51
Aerobacter aerogenes 13048	14.3	53	17.9	45
Klebsiella pneumoniae	5.7	21	10.2	25
Proteus vulgaris	3.0	11	5.5	14
Salmonella typhimurium	23.5	87	27.9	71
Serratia marcescens 4180	2.1	8	2.8	7
Serratia marcescens S.M. 11	1.6	6	...	...
Shigella dysenteriae	23.8	88	27.7	71
Aeromonas hydrophila	1.2	4	...	...
Bacillus subtilis	0.4	1	...	...
Pseudomonas aeruginosa	0.5	2	0.4	1
T2 bacteriophage	0.3	1	0.4	1
Calf thymus	0.4	1	0.5	1
Mouse liver	0.4	1	...	...

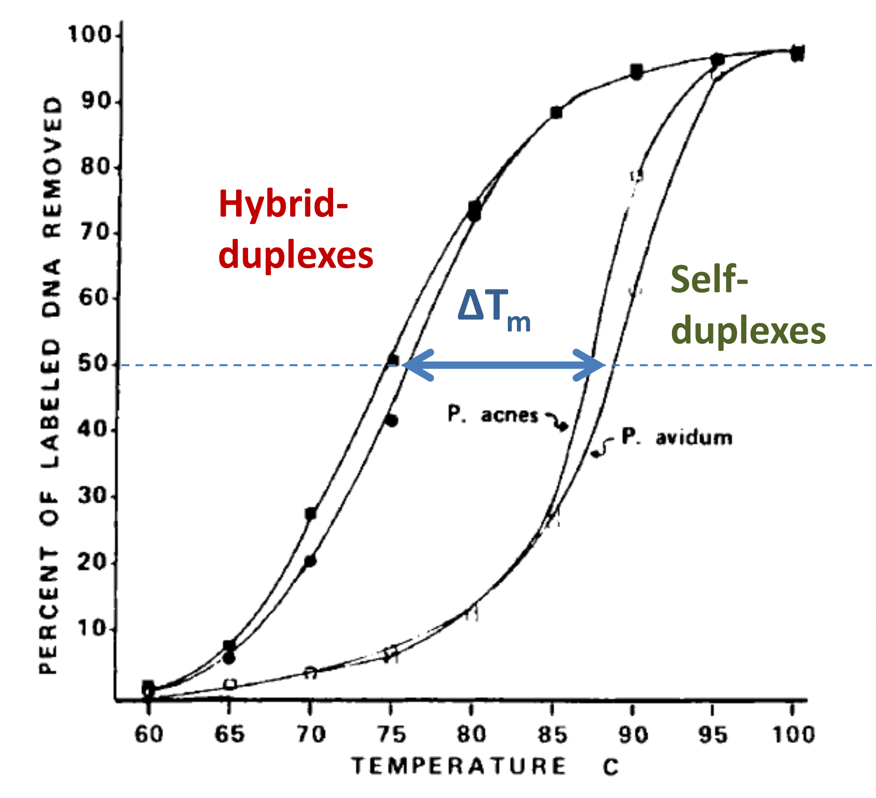
In the left-hand columns are given the results of experiments in which 50 μg of *E. coli* pulse-labeled RNA was incubated with 0.5 gm of the various DNA-agar preparations. Where *E. coli* sheared, denatured DNA was used (right-hand columns), 15 μg was incubated with a quantity of agar containing 150 μg of trapped DNA (about 0.5 gm).

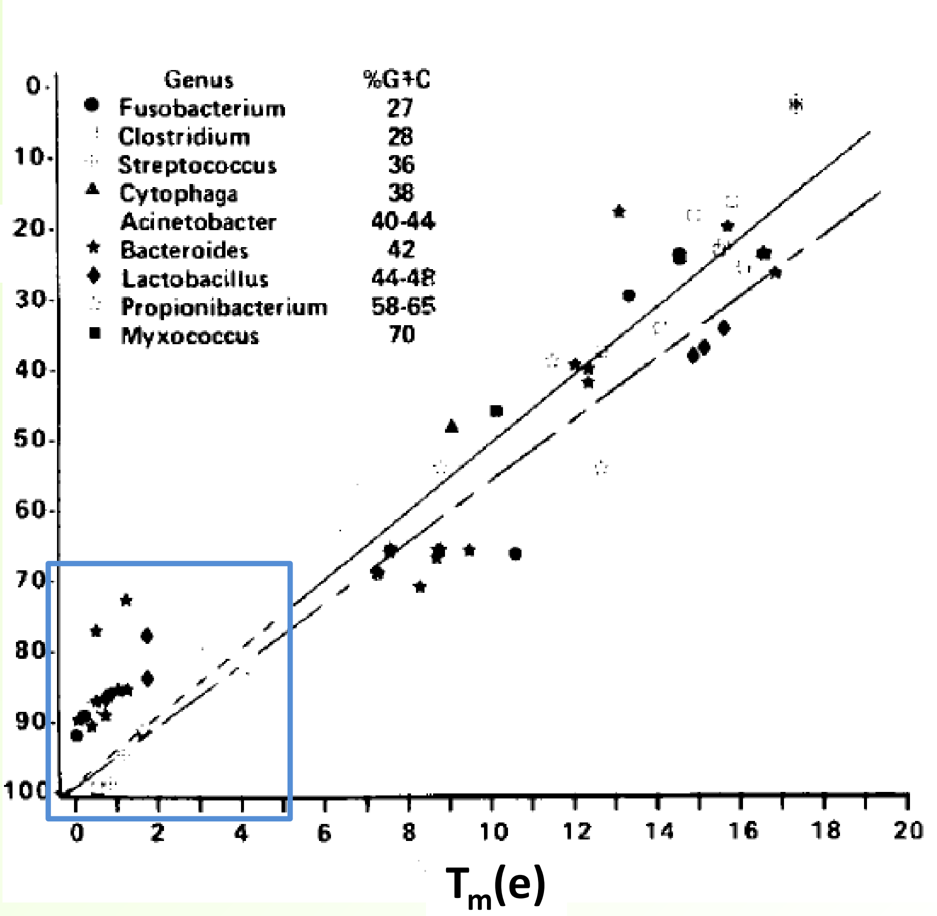
McCarthy, B.J. and Bolton, E.T. (1963) An approach to the measurement of genetic relatedness among organisms., *Proc. Natl. Acad. Sci. U.S.A.* **50** (1) 156-164

Another hybridisation metric: $\Delta \boldsymbol{T_m}$

- Self versus self DNA reassociates to form duplexes
- DNA from two different organisms associates (hybridises) to form hybrid duplexes
- How stable are these?
  - At what temperature has 50% of Self duplex dissociated?
  - At what temperature has 50% of Hybrid duplex dissociated?
  - The difference between these is **ΔT_m** (or **T_m(e)**)

Johnson (1973)





INTRA-SPECIES: Almost all of the data points for **two strains classified as the same species** are in the blue box

INTER-SPECIES: NONE of the inter-species strain associations are in the box

What the *Ad Hoc Committee* said…

- “At present, the species is the only taxonomic unit that can be defined in phylogenetic terms.
- *The **phylogenetic definition of a species** generally would include strains with approximately **70% or greater [relative DNA-binding*]** and with **5°C or less ΔT_m**. Both values must be considered.*
- Phenotypic characteristics should agree with this definition and would be allowed to override the phylo-genetic concept of species only in a few exceptional cases.”
- - Wayne *et al.* (1987) Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics, , *Int. J. Syst. Bacteriol.*, **27** (1) 44-57

- * The literal term used was “*DNA-DNA relatedness*” – it’s a terminology matter; DO NOT CONFUSE THIS WITH % DNA SEQUENCE SIMILARITY! Which is why it’s been replaced with “% relative DNA-binding” here

So how does **DNA sequence identity** correlate with this?

- Unsurprisingly, the amount of DNA-reassociation depends on the number of cognate base pairs versus base mispairs
- Studies (1970s onwards) examined mispairs in oligonucleotides →measurable **sequence identity**
  - indicated that thermal stabilities decrease by ~ 1 to 2% for each percent of the genomic DNA which mispairs
  - (see Stackebrandt & Goebel, 1994)
  - no measurable reassociation unless pairing is ≥ **85%**
  - ≥ **70%** relative DNA association → **96%** sequence identity
  - That's **whole-genome DNA**

- ≥ **70%** relative DNA association

  → **97%** sequence identity of 16S rRNA gene
  - (remember that's the whole gene)
- **But the converse does not hold**
- There are plenty of known cases of pairs of bacteria with:

  ≥ 97% 16S rRNA identity but **< 70%** relative DNA binding (some <<<< 70%)
- What this basically means is:
  - if a pair of prokaryotes have **< 97%** 16S rRNA sequence identity
  - then they are **not** members of the same species*
  - if they have **≥ 97%** 16S rRNA sequence identity
  - then they **might be**; but they **might not be**
  - (and remember to check **T_m(e)** as well)
- * …or < **98.7%** identity, depending on whom you agree with

That’s dealing with full-length 16S rRNA genes

Maybe the situation is better with the regions we amplify?

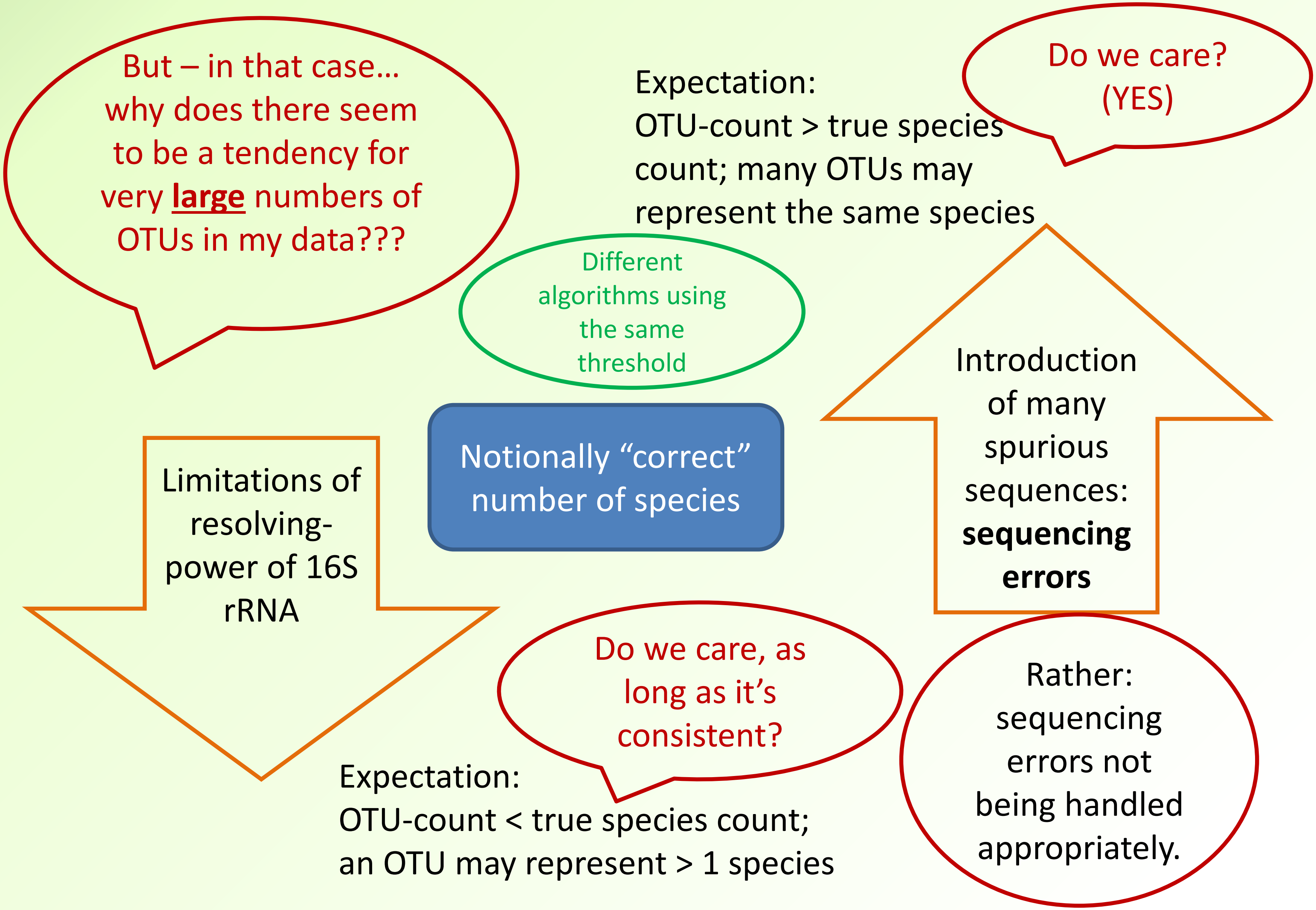
(No, it’s worse)

Various degrees of [sequence identity] in stretches of 200 nucleotides along the primary structure of pairs of 16S rRNAs from organisms with different degrees of relatedness (after Stackebrandt & Goebel, 1994)

Position	**16S rRNA sequence identity (%) between:**		
	Streptomyces ambofaciens and *Streptomyces violaceoruber*	*Mycobacterium phlei* and *Mycobacterium tuberculosis*	*Aeromicrobium erythreum* and *Rhodococcus fascians*
Overall	98.8	96.4	90.9
0-200	**96.3**	94.1	**80.7**
201-400	98.4	97.8	94.6
401-600	100.0	93.1	94.6
601-800	99.0	97.9	85.7
801-1000	100.0	100.0	94.0
1001-1200	98.9	**92.8**	90.0
1201-1400	99.5	100.0	94.0

Approx. position of V4-V5 amplicons

- “evidence is strong that sequence analyses of 16S rRNA is not the appropriate method to replace DNA reassociation for the **delineation of species** and measurement of **intraspecies relationships**”
  - – Stackebrandt & Goebel (1994)
- This is all another way of saying…
  - – If you are sequencing 16S rRNA gene amplicons (even if they were full-length), don’t expect to resolve a microbiome to finer than genus level
  - – Never mind the differences between strains



But – in that case… why does there seem to be a tendency for very **large** numbers of OTUs in my data???

Expectation:
OTU-count > true species count; many OTUs may represent the same species

Do we care?
(YES)

Different algorithms using the same threshold

Introduction of many spurious sequences: **sequencing errors**

Notionally "correct" number of species

Limitations of resolving-power of 16S rRNA

Do we care, as long as it's consistent?

Expectation:
OTU-count < true species count; an OTU may represent > 1 species

Rather: sequencing errors not being handled appropriately.

So what’s happened in the last 20+ years?

Learning to love backward compatability ☺

Recent years

- E.g. 2010:
- “Given the considerable promise whole-genome sequencing offers for phylogeny and classification, it is surprising that microbial systematics and genomics have not yet been reconciled.”
  - Klenk & Göker (2010) En route to a genome-based classification of Archaea and Bacteria?, Syst. Appl. Microbiol. **33** (4) 175-182

- E.g. 2013: (Meier-Kolthoff, Auch, Klenk & Göker, Genome sequence-based species delimitation with confidence intervals and improved distance functions, BMC Bioinformatics 14:60):
  - In essence, about **computational methods for predicting DNA-DNA Hybridisation (DDH) from genome sequences**

Backward-compatability

- Meier-Kolthoff *et al.* (2013):
  - “If the genomic DNA of two respective organisms reveals a DDH **[DNA-DNA-Hybridisation] similarity of below 70% this is the main argument to regard them as distinct species**…”
  - “The increasing availability of genome sequences thus triggered **the development of computational techniques to replace wet-lab DDH….**
  - “**unless high correlations with wet-lab DDH**, and precise models for estimating DDH or at least DDH-analogous species boundaries from genome-to-genome comparisons, were available, **the newly calculated values were not comparable to the previous ones** and could yield largely deviating species-boundary estimates and, thus, **an inconsistent microbial taxonomic classification.**
  - Hence, for obvious reasons the literature on in-silico replacements for DDH considered **correspondence with wet-lab DDH values as optimality criterion**. As a consequence, regression and/or correlation analyses with wet-lab DDH values were used throughout for the calibration and optimization of the in-silico replacement methods”

Polyphasic taxonomy
(still)
rules OK!