Introducing Microbiome Bioinformatics

Part 2.

Recap: Aims

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

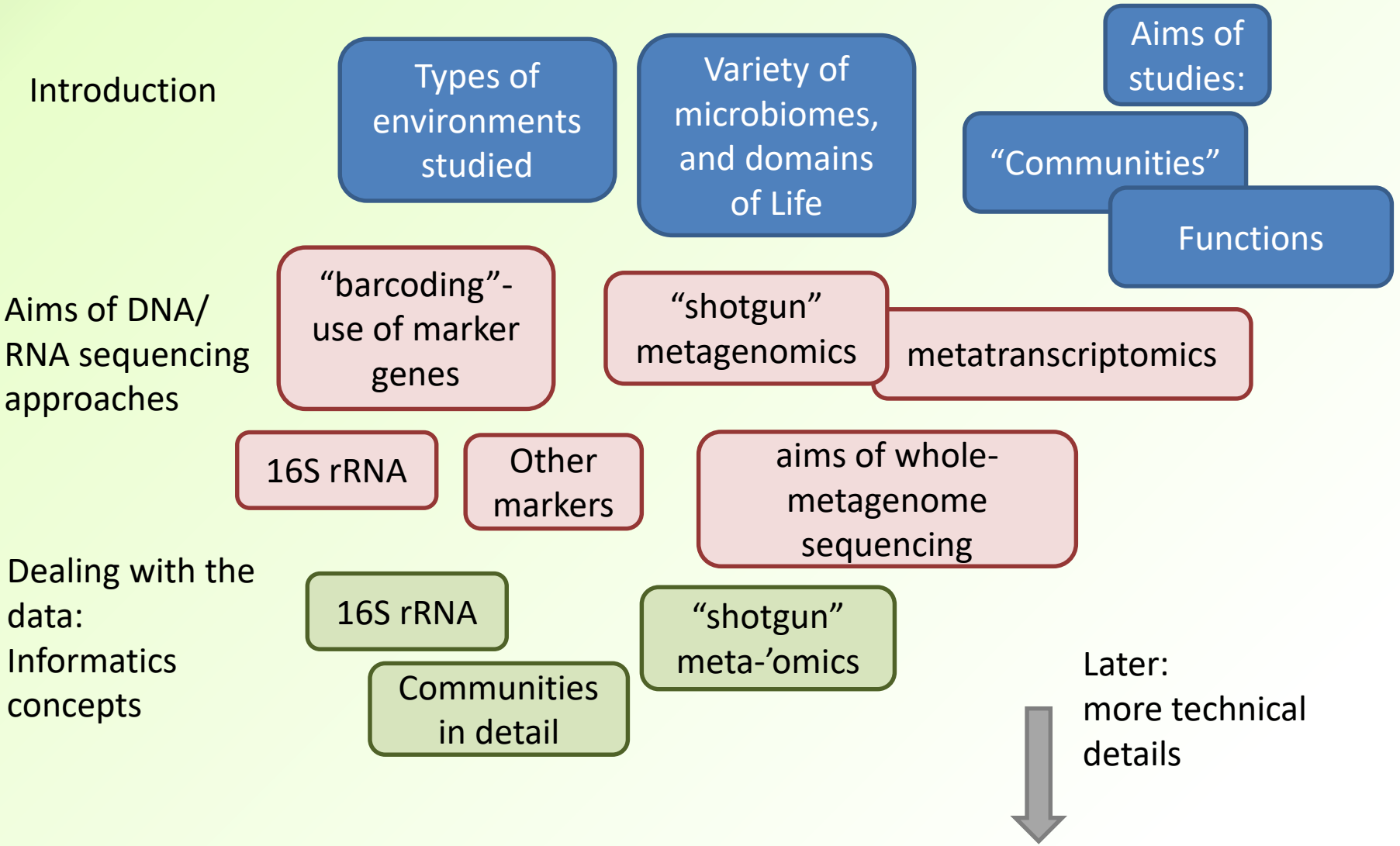
Series of talks

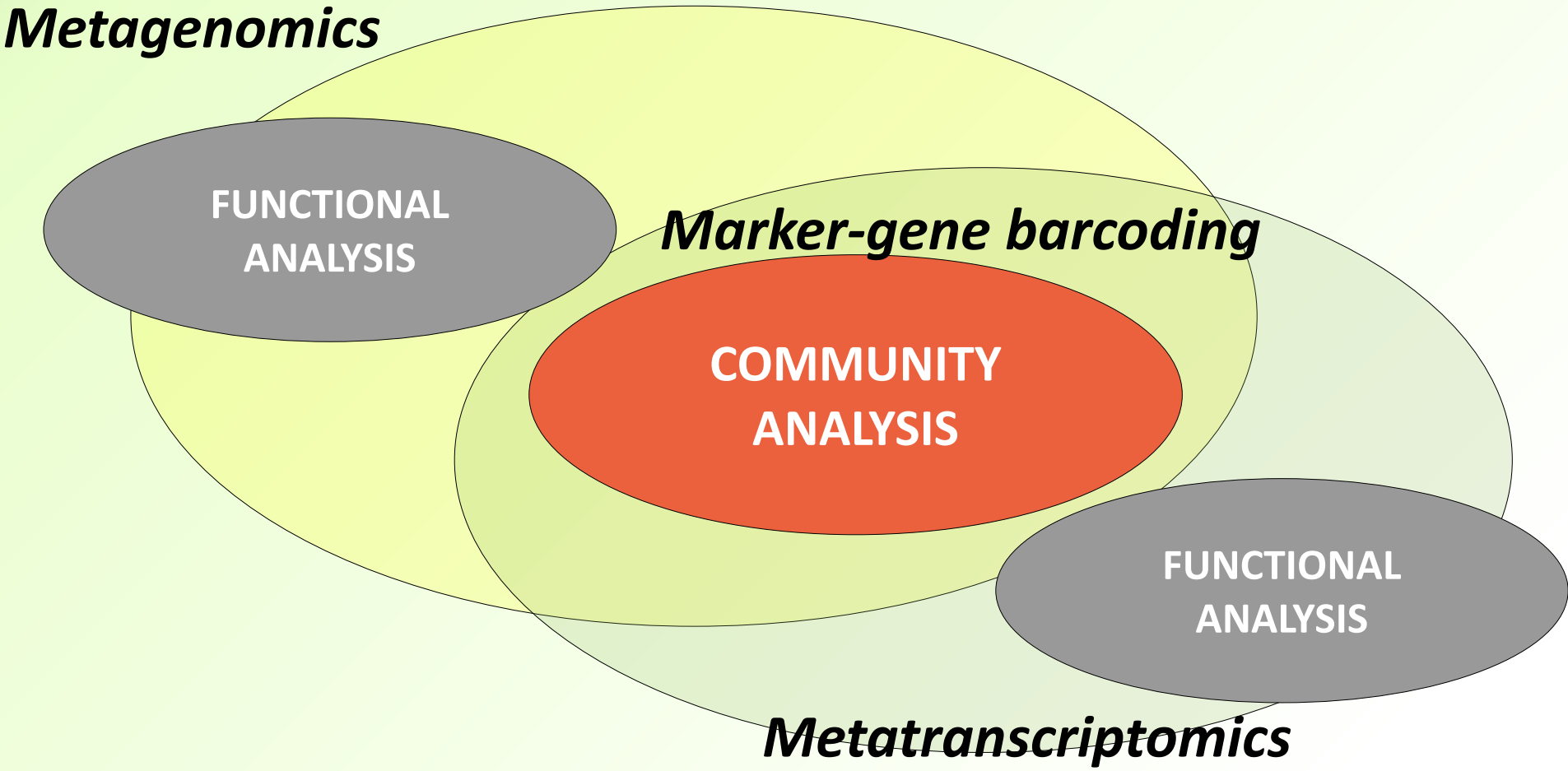
- At least 3 sessions to cover what I would like
- Beyond that – if there is demand –
  - can progress to more technical talks
  - especially about 16S analysis (probably)
  - increasingly metagenomics in GHFS research
- Informal and flexible
  - Please interrupt and ask questions
  - Suggestions for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

Series of talks

- Part 1: 27/1/2017
  - "Biological and Experimental Stuff that a microbiome bioinformatician needs to know"
  - http://ghfs1.ifr.ac.uk/ghfs/
  - (see post of the above date)
- Part 3: 24/2/2017, 14:00, Barton

Topics, top-down




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

More on terminology…

"…The approach involves directly accessing the genomes of soil organisms that cannot be … cultured by isolating their DNA….
The methodology has been made possible by advances in molecular biology … which have laid the groundwork for cloning and functional analysis of **the collective genomes of soil microflora, which we term the metagenome of the soil."**
Handelsman *et al.* (1998) *Chemistry and Biology* **5** R245-9

"…. This study provides positive validation of the effectiveness of **targeting 16S metagenomes using short-read sequencing technology**."
Sundquist *et al.* (2007) Bacterial flora-typing with targeted, chip-based Pyrosequencing *BMC Microbiology* **7** 108

Tikhonov *et al. (*2015) Interpreting **16S metagenomic data** without clustering to achieve sub-OTU resolution
ISME Journal **9** (1) 68-80

Definitions used here

- Best not to refer to 16S studies as “metagenomics”
- This will be referred to here as “16S-barcoding” or “16S-based community analysis”
  - Not to be confused with “barcoding” in the sequencing platform context
- Alternative names for 16S studies have been proposed, e.g. “taxonomics”

(Shotgun) Metagenomics "whole-metagenome sequencing"

"Sequence everything"

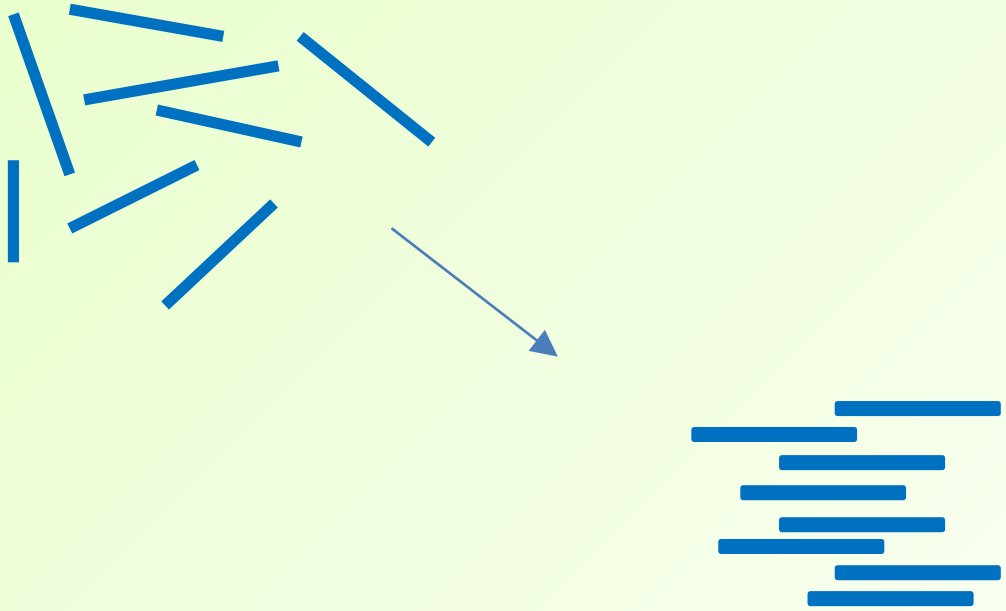
- *Everything you can*
- *Everything you are interested in*

- *Who* : phylotypes/taxa – same goals as marker gene amplicons
- *What* : genes → potential functions → potential pathways
- *Beyond the census* : discovering and assembling new genome sequences? (maybe)
- *Not-whole* : unless it's a very narrow community, you are **sampling,** not fully sequencing the "whole metagenome"

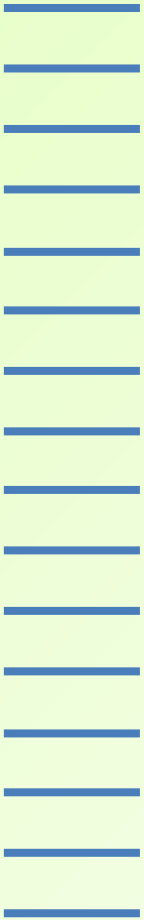
Shotgun metagenomics

- Goal: randomly sample the genomic sequences of any organisms present. So, randomly samples genes, regulatory regions,etc
- These can be compared with database sequences
  - If the database sequences are annotated with:
    - Functional information,
      - then you can do functional profiling
    - Phylotypic/Taxonomic information
      - then can do community profiling

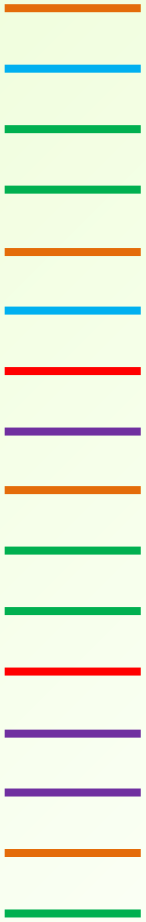
What about “assembly”?



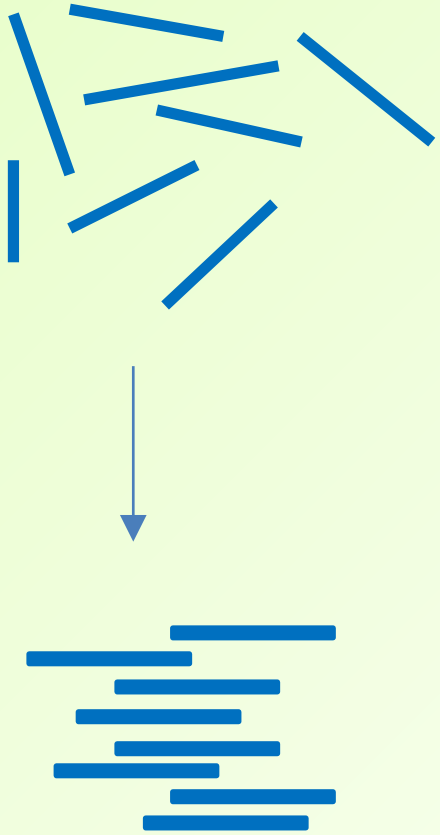
Contrast with 16S

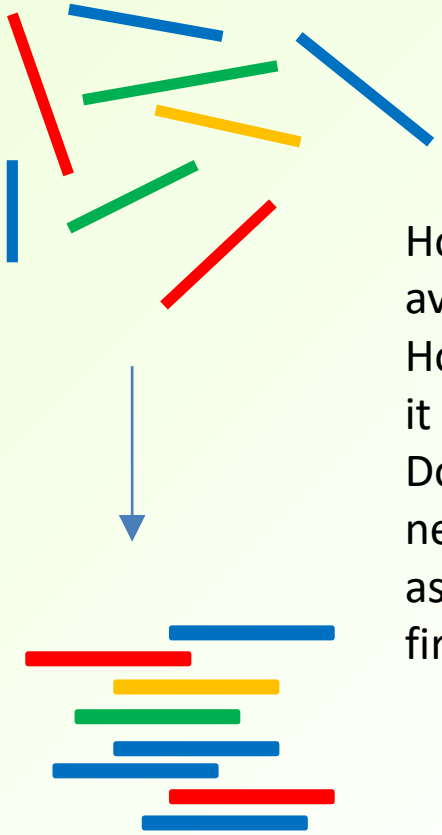


Sequencing lots of different organisms' version of the same thing (marker gene)



Assembly or clustering? Chimaeras?



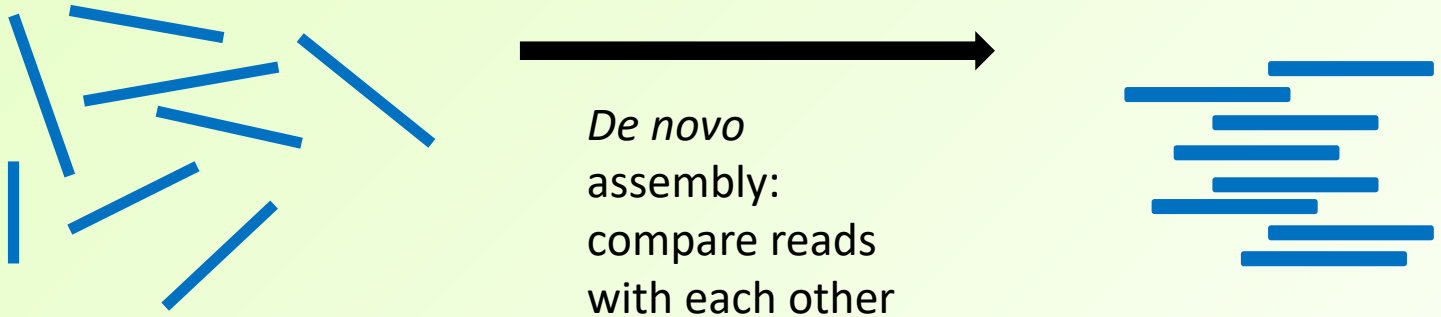


How much is this avoidable?
How much does it matter?
Do you even need to attempt assembly in the first place?

Metagenome read “assembly”

- What about “assembling” your metagenomic reads?
  - **Can you do it?** To a partial extent, if you have a sufficient number of reads. Software tools available. Also depends on **community structure**
  - **Do you really need to do it?**
    - Sometimes **yes**, it’s very useful, in detailed studies
      - short reads -> longer gene sequences / genome fragments
      - Longer sequences can **greatly reduce ambiguities** in matching reads to database sequences
      - **Helps in particular with identifying organisms**
      - also can help with identifying types of **genes**
    - *May not be necessary to attempt this*
    - **Consider what the assembled genomes will tell you which the sets of unassembled reads will not**
  - Nowadays, it can be relatively “straightforward” to do some sort of assembly/clustering; can be computationally intensive

Assembly using a reference genome



De novo assembly: compare reads with each other

"Resequencing" : align reads with reference genome sequence

- Clearly, reference genomes help with identifying/assembling metagenomic reads
- Does not solve all problems
- May still be ambiguities
- A lot depends on what sequences are available in the databases

Shotgun reads → whole genomes? → Whole metagenomes?

- Assembly of some reads to at least a partial extent can be very helpful
- Can you extract/assemble on or more **entire genome** sequence(s) from a set of more diverse metagenomics reads?
  - Yes, sometimes
- Can you sequence/assemble **all** of the genomes of the organisms present?
  - If the nature of the community is tractable, then yes
  - Some fairly early studies came close to achieving this

A low-diversity biofilm

- Metagenomics of acid mine drainage biofilm
- Tyson *et al.* (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment, *Nature* **428** 37-43
- Sanger/capillary-sequencing
  - Longer reads, but much lower read number than later platforms
- Biofilm dominated by 5 species
  - Obtained 2 near-complete genomes (still quite fragmented)
- "using random shotgun sequencing of DNA from a natural acidophilic biofilm, we report reconstruction of near-complete genomes of *Leptospirillum* group II and *Ferroplasma* type II, and partial recovery of three other genomes.
- This was possible because the biofilm was dominated by a small number of species populations and the frequency of genomic rearrangements and gene insertions or deletions was relatively low.
- "The *Ferroplasma* type II genome seems to be a composite from three ancestral strains that have undergone homologous recombination to form a large population of mosaic genomes."

Untangling Genomes from Metagenomes: Revealing an Uncultured Class of Marine Euryarchaeota

Vaughn Iverson, Robert M. Morris, Christian D. Frazar, Chris T. Berthiaume, Rhonda L. Morales, E. Virginia Armbrust*

+ Author Affiliations

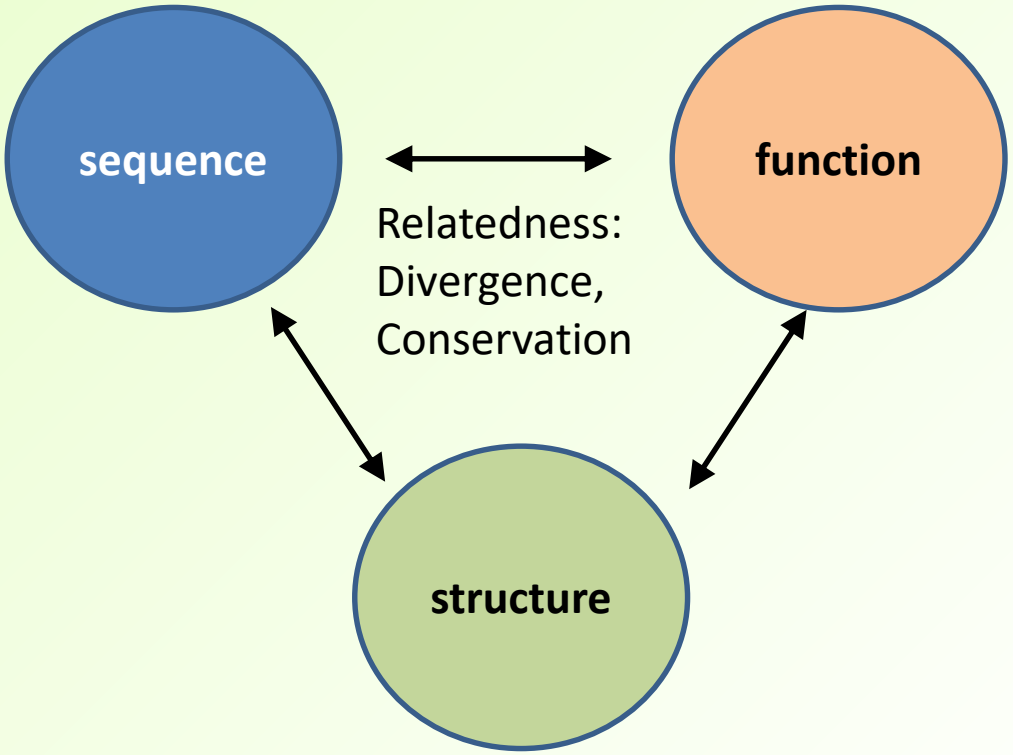
*To whom correspondence should be addressed. E-mail: armbrust@uw.edu

Science 03 Feb 2012:
Vol. 335, Issue 6068, pp. 587-590
DOI: 10.1126/science.1212665

Abstract

Ecosystems are shaped by complex communities of mostly unculturable microbes. Metagenomes provide a fragmented view of such communities, but the ecosystem functions of major groups of organisms remain mysterious. To better characterize members of these communities, we developed methods to reconstruct genomes directly from mate-paired short-read metagenomes. We closed a genome representing the as-yet uncultured marine group II *Euryarchaeota*, assembled de novo from 1.7% of a metagenome sequenced from surface seawater. The genome describes a motile, photo-heterotrophic cell focused on degradation of protein and lipids and clarifies the origin of proteorhodopsin. It also demonstrates that high-coverage mate-paired sequence can overcome assembly difficulties caused by interstrain variation in complex microbial communities, enabling inference of ecosystem functions for uncultured members.

The *what* question (function)



Conservation versus divergence: the *what* and *who*

Gene sequences can diverge a lot, with function still conserved

Very good news if your metagenomics reads contain novel organisms
Also means that chimaeric "assemblies" don't matter much

Gene sequences of fairly closely-related organisms can be quite diverged

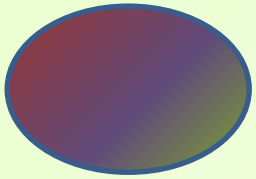
But conversely…

Gene sequences of distantly-related organisms can be very conserved

It depends on the gene

Thinking at functional level
means thinking largely at the protein level
(ignoring noncoding RNAs for the moment)

Gene/protein sequences can diverge a lot, with function still conserved



XYZase

MAEDLLSQTNNI.....

Your metagenomics reads code for a protein which is…

95% identical by sequence

So you can be sure that your metagenome-coded protein is also an XYZase

- even with considerably lower sequence identity
- may be a closely-related variant function

75%

60%

….

No “rules” which say precisely how low a % identity this holds for

Caveats…

Gene/protein families

New | Top twenty | Numbers | A | B | **C** | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z

ID	Accession	Number of sequences		Average length	Average %id	Average coverage	Has 3D	Change status	Description
		Seed	Full						
C-C_Bond_Lyase	PF15617	68	232	317.70	41	89.34		Changed	C-C_Bond_Lyase of the TIM-Barrel fold
c-SKI_SMAD_bind	PF08782	42	412	91.60	46	14.13	✓	Changed	c-SKI Smad4 binding domain
C1-set	PF07654	71	5062	84.20	25	34.13	✓	Changed	Immunoglobulin C1-set domain
C1-set_C	PF16196	30	291	50.50	47	20.32	✓	Changed	C1-set C-terminal domain
C1q	PF00386	32	2794	123.60	28	33.59	✓	Changed	C1q domain
C1_1	PF00130	44	8266	52.70	30	7.86	✓	Changed	Phorbol esters/diacylglycerol binding domain (C1 domain)
C1_2	PF03107	220	5454	48.40	30	31.83	✓	Changed	C1 domain
C1_4	PF07975	11	440	55.20	42	12.14	✓	Changed	TFIIH C1-like domain
C2	PF00168	261	39609	106.00	18	20.55	✓	Changed	C2 domain
C2-C2_1	PF11618	49	242	138.60	32	12.11	✓	Changed	First C2 domain of RPGR-interacting protein 1
C2-set	PF05790	27	218	81.00	27	25.37	✓	Changed	Immunoglobulin C2-set domain

Pfam protein families database (pfam.xfam.org)

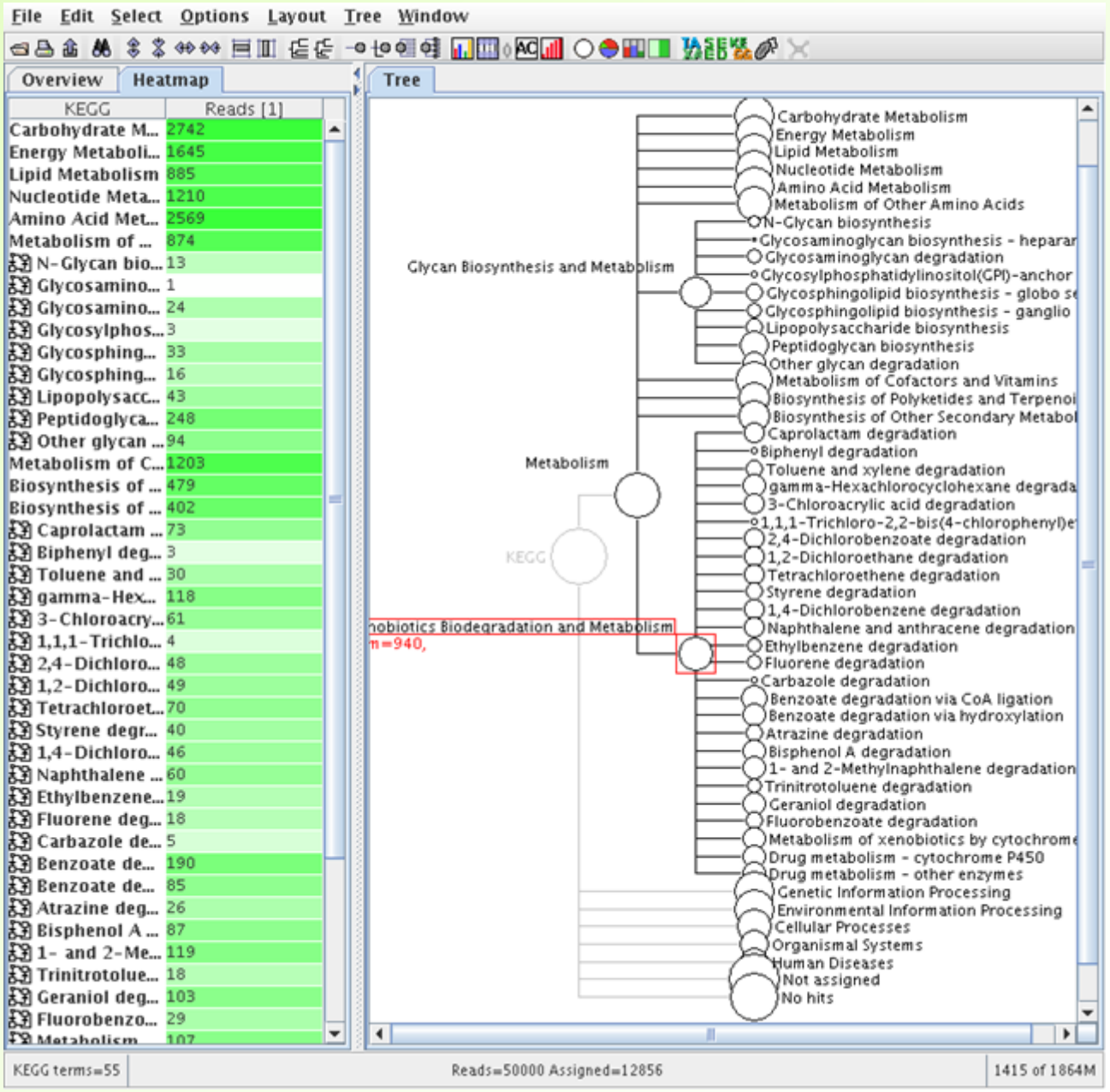
The bad side of the *what* question

- So it's easy to assign function to all of your reads? No..,.
- Some of your reads might be from a completely novel gene
  - Nothing close to it is in any databases. So you can't identify it.
  - This is also bad news for the "*who*" question

The bad side of the *what* question

- More commonly –
- Another bunch of your reads are clearly similar to numerous gene sequences in the databases
  - Some of these are even in sequenced genomes
    - so that answers the "*who*" question
  - **But, nobody knows what these genes actually do**

- "*E. coli* K-12 and yeast *Saccharomyces cerevisiae* appear to be the only organisms for which at least 50% of the genes have been studied experimentally"
  - Galperin & Koonin (2010) From complete genome sequence to "complete" understanding? *Trends Biotechnol.* **28**(8) 398-406
- Predicting function: The "70% hurdle"


File Edit Select Options Layout Tree Window
Overview
Heatmap
KEGG
Reads [1]
Carbohydrate M... 2742
Energy Metaboli... 1645
Lipid Metabolism 885
Nucleotide Meta... 1210
Amino Acid Met... 2569
Metabolism of ... 874
N-Glycan bio... 13
Glycosamino... 1
Glycosamino... 24
Glycosylphos... 3
Glycosphing... 33
Glycosphing... 16
Lipopolysacc... 43
Peptidoglyca... 248
Other glycan ... 94
Metabolism of C... 1203
Biosynthesis of ... 479
Biosynthesis of ... 402
Caprolactam ... 73
Biphenyl deg... 3
Toluene and ... 30
gamma-Hex... 118
3-Chloroacry... 61
1,1,1-Trichlo... 4
2,4-Dichloro... 48
1,2-Dichloro... 49
Tetrachloroet... 70
Styrene degr... 40
1,4-Dichloro... 46
Naphthalene ... 60
Ethylbenzene... 19
Fluorene deg... 18
Carbazole de... 5
Benzoate de... 190
Benzoate de... 85
Atrazine deg... 26
Bisphenol A ... 87
1- and 2-Me... 119
Trinitrotolue... 18
Geraniol deg... 103
Fluorobenzo... 29
Metabolism 107
Tree
Carbohydrate Metabolism
Energy Metabolism
Lipid Metabolism
Nucleotide Metabolism
Amino Acid Metabolism
Metabolism of Other Amino Acids
N-Glycan biosynthesis
Glycosaminoglycan biosynthesis - heparar
Glycosaminoglycan degradation
Glycosylphosphatidylinositol(GPI)-anchor
Glycosphingolipid biosynthesis - globo se
Glycosphingolipid biosynthesis - ganglio
Lipopolysaccharide biosynthesis
Peptidoglycan biosynthesis
Other glycan degradation
Glycan Biosynthesis and Metabolism
Metabolism of Cofactors and Vitamins
Biosynthesis of Polyketides and Terpenoi
Biosynthesis of Other Secondary Metabol
Caprolactam degradation
Biphenyl degradation
Toluene and xylene degradation
gamma-Hexachlorocyclohexane degrada
3-Chloroacrylic acid degradation
1,1,1-Trichloro-2,2-bis(4-chlorophenyl)e
2,4-Dichlorobenzoate degradation
1,2-Dichloroethane degradation
Tetrachloroethene degradation
Styrene degradation
1,4-Dichlorobenzene degradation
Metabolism
KEGG
nobiotics Biodegradation and Metabolism
m=940,
Naphthalene and anthracene degradation
Ethylbenzene degradation
Fluorene degradation
Carbazole degradation
Benzoate degradation via CoA ligation
Benzoate degradation via hydroxylation
Atrazine degradation
Bisphenol A degradation
1- and 2-Methylnaphthalene degradation
Trinitrotoluene degradation
Geraniol degradation
Fluorobenzoate degradation
Metabolism of xenobiotics by cytochrome
Drug metabolism - cytochrome P450
Drug metabolism - other enzymes
Genetic Information Processing
Environmental Information Processing
Cellular Processes
Organismal Systems
Human Diseases
Not assigned
No hits
KEGG terms=55
Reads=50000 Assigned=12856
1415 of 1864M

Metagenomics and the "*who*" question

A bit about bias

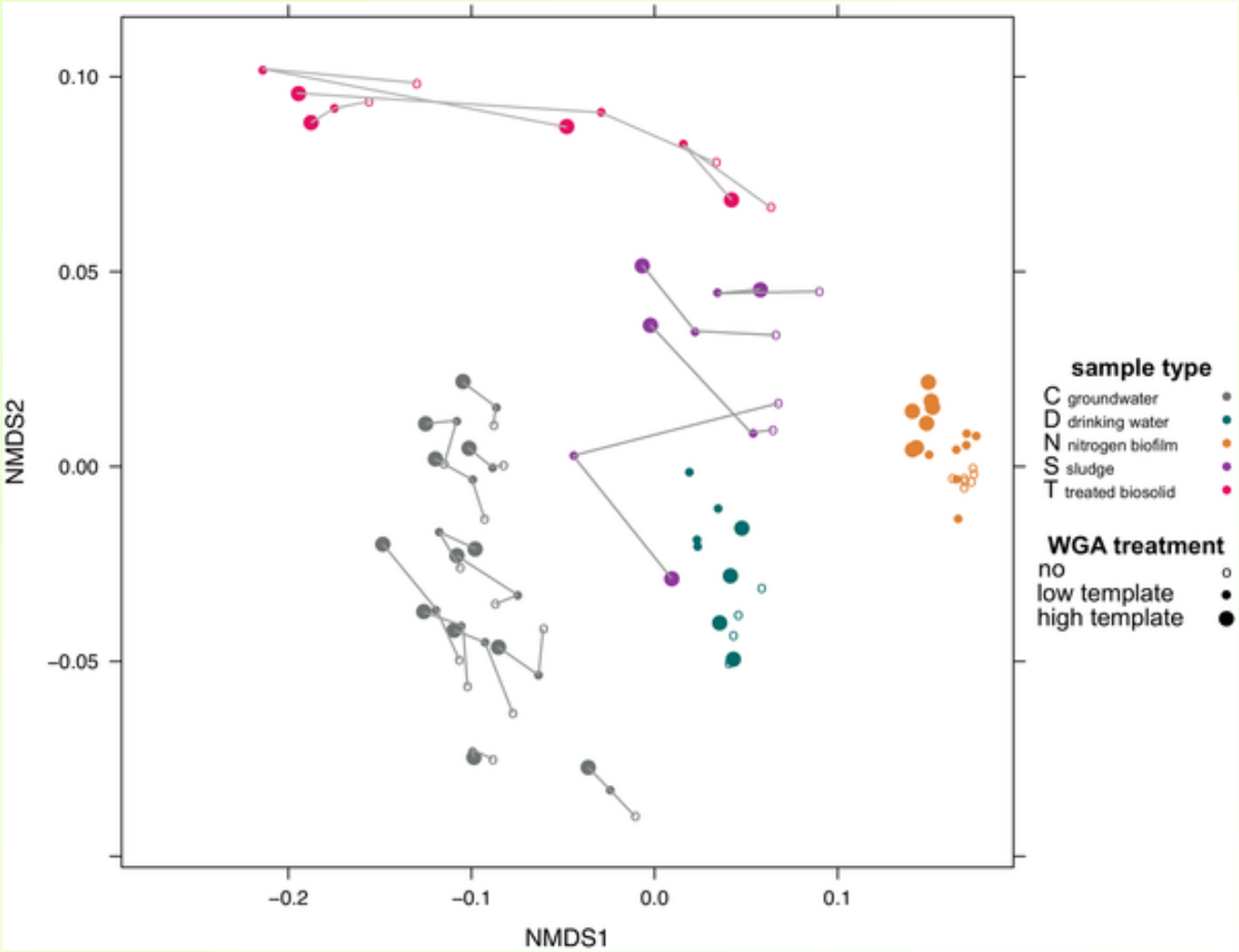
How do you get your DNA to put in the sequencer?

- Contrast to PCR-amplification of marker genes, which selects particular taxa, such as:
  - Archaea and/or Bacteria; or Fungi
- You may be interested in particular organisms
  - Which informs the DNA-extraction process
    - E.g. viromes from ocean or gut
    - Removal of unwanted 'extras' can be done at the post-sequencing stage, *in silico*
- DNA-amplification is often necessary
  - Which should amplify all organisms' DNA equally
  - But almost certainly won't

“Whole-genome amplification”

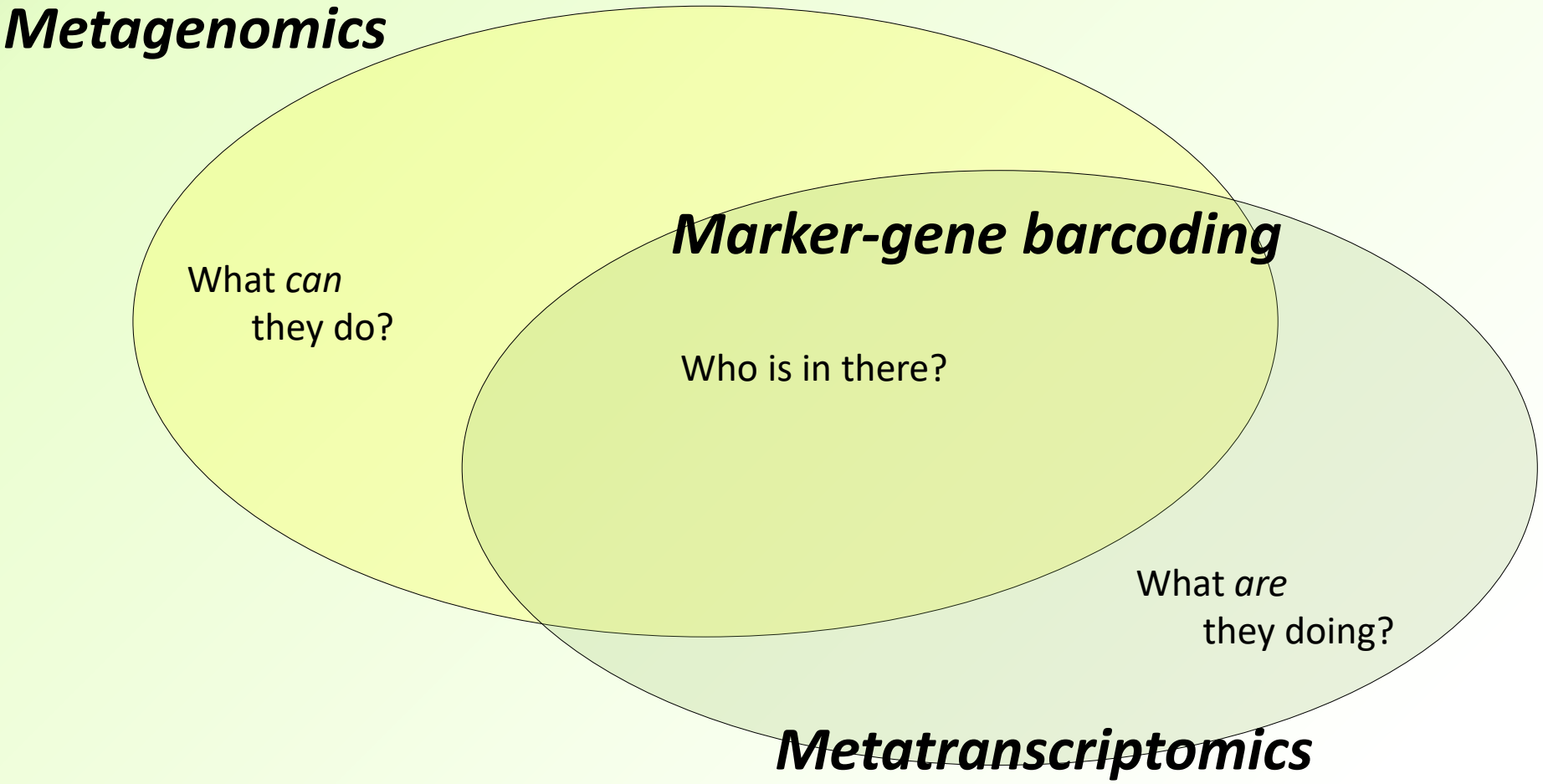
- Used in traditional genome sequencing
- Also in a single-cell context
- And in metagenomics
- The randomness of the sampling is based on using random hexamer primers
  - A.k.a. Multiple Displacement Amplification
  - Uses DNA polymerase from Φ29, a *Bacillus* phage
- A number of studies have found that it biases in favour of low-GC-content genomes
  - Biased against many Actinobacteria; some α- and β-Proteobacteria
  - The bias is not always uniform even for one type of microbiome
- Whether there is an exact match with the primer sequence, has been found to be less important
- In some studies the effect of the bias has been shown to be a lot smaller than the differences between microbiomes

Fig 1. Various effects of WGA treatment of samples from five biotopes investigated by ordination analysis.

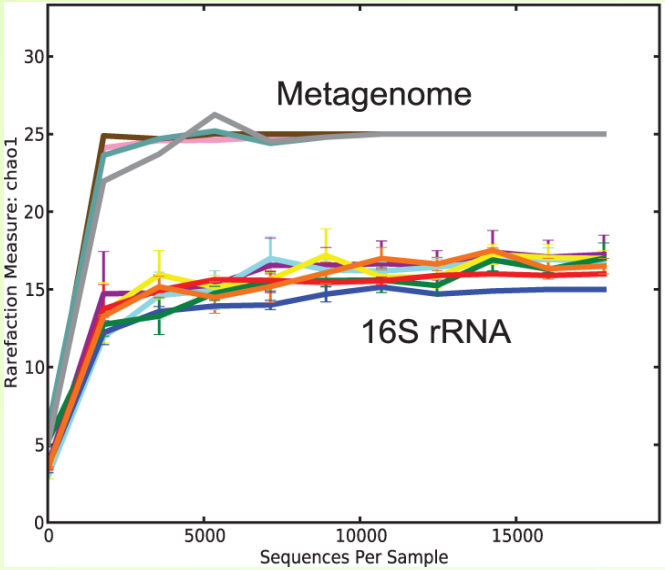


Probst AJ, Weinmaier T, DeSantis TZ, Santo Domingo JW, Ashbolt N (2015) New Perspectives on Microbial Community Distortion after Whole-Genome Amplification. PLOS ONE 10(5): e0124158. doi:10.1371/journal.pone.0124158
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0124158

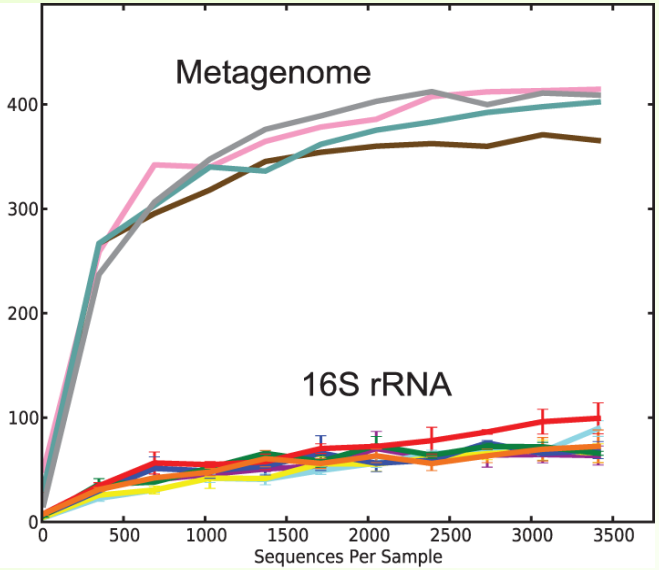
So, do marker-gene sequencing and metagenomic sequencing give us the same answers?



Estimated no. of **phyla**



Estimated no. of **genera**



Poretsky R, Rodriguez-R LM, Luo C, Tsementzi D, et al. (2014) Strengths and Limitations of 16S rRNA Gene Amplicon Sequencing in Revealing Temporal Microbial Community Dynamics. PLoS ONE 9(4): e93827. doi:10.1371/journal.pone.0093827

Example anomalies: *Thiomonas*

- 16S says 45% of the reads are *Thiomonas*
- 16S says 23% of the OTUs are
- WGS-metagenomic contigs says 0.3% are
- WGS-single reads:
  - 10% of 16S gene fragments are

- > 200 *Thiomonas* species/OTUs in the 16S database
- But only 2 *Thiomonas* reference genomes

- US-UK Bioinformatics for the Microbiome Workshop
- "**Summary of findings**
  - **Standards are necessary to move forward.** Validating sequencing, metabolomics, and culture-based pipelines are imperative. **At the moment, the field is considered by some to be a "pre-science"** because labs are often not able to reproduce each other's results. Some argued that instead of labelling the research a "pre-science," maybe there are missing variables. Guidelines for collecting and storing samples are essential, as well as high-quality reference materials.
  - **Informatics resources are vital for breakthroughs in microbiome research.** At the moment, resources are incomplete, especially in regards to metagenomes, metabolites, metadata and gene catalogs. …."