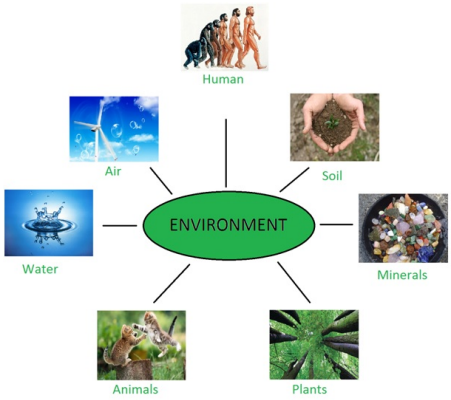


METAGENOMICS

Omics Techniques

Bàrbara Negre, UAB

METAGENOMICS: Why should we care?

(1) Almost all **antibiotics and anticancer drugs** are naturally occuring metabolites produced by microbes. Without finding new microbes, we won't find new drugs.

(2) Microbes in our bodies cause disease, **prevent disease**, and interact with our diet and genome to modify their effect on health. Without knowing the makeup of each person's internal microbiome, our ability to diagnose and cure diseases will always be limited.

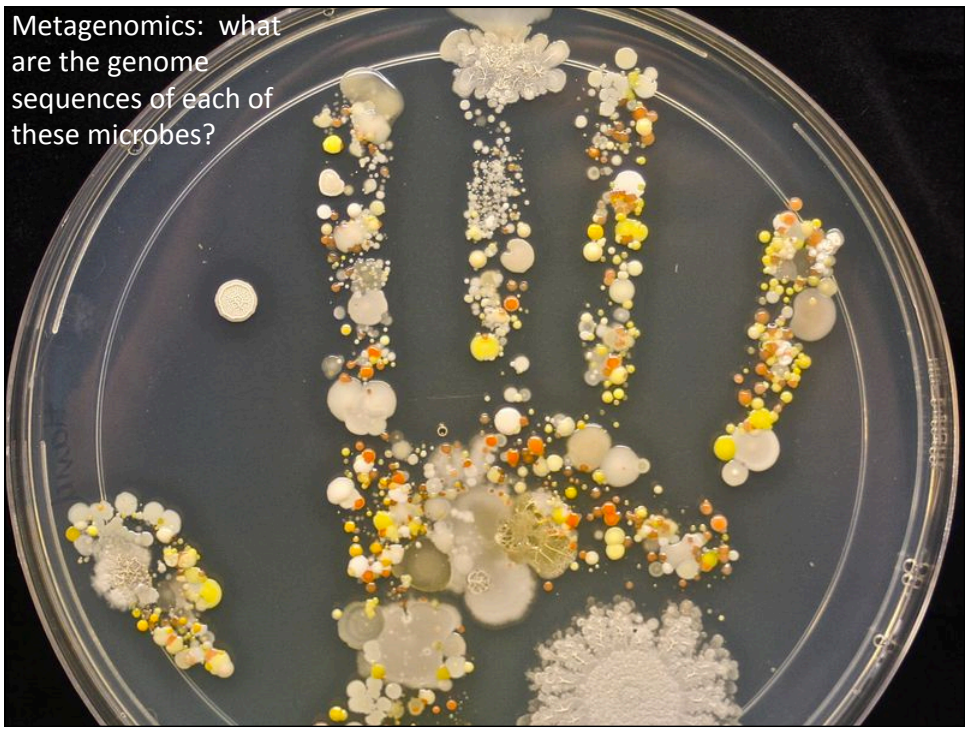
(3) Microbes are found inside of many tumors, and **treatment of tumors** with antibiotics (which do not affect human cells) can either increase or decrease the growth rate of the tumor: microbes play strange and poorly understood roles in many diseases that we wouldn't normally think of as being microbial in nature.

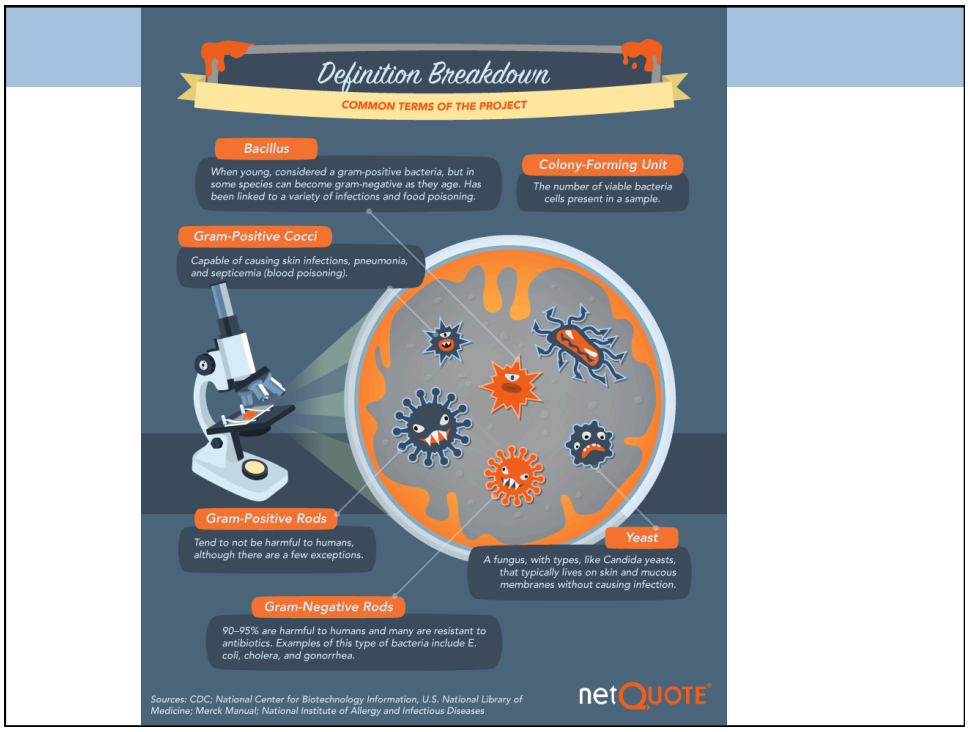
(4) The best way to **understand evolution** is to obtain sequences for thousands of genomes.

(5) Microbes found in contaminated waters (eg: pig shit in farms, heavy-metal contamination from old factories) can often eat toxic compounds, and turn them into non-toxic compounds, providing a possible solution to pollution. **Microbes as cleaners**.

Metagenomics: How do we do it?

What microbes can we find where?





CFUs in different environments

Why do toilet seats have such low CFUs?

extra credit if you can find me CFU data for various surfaces in Barcelona (eg: metro, beaches, river)

The coffee-machine bacteriome

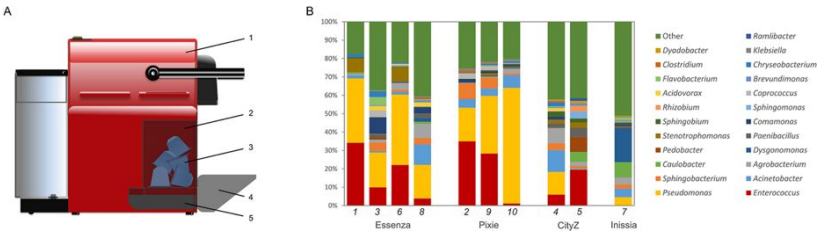


Figure 1. (A) Schematic representation of a Nespresso machine (1) including a capsule (3) container (2), cup tray (4) and a drip tray (5), which was sampled in this work. (B) Bacterial profile of the nine Nespresso machines sampled according to 16S rRNA gene sequencing. Samples numbered in accordance to Table 1.

SCIENTIFIC REPORTS

OPEN

Received: 12 June 2015
Accepted: 16 October 2015
Published: 23 November 2015

The coffee-machine bacteriome: biodiversity and colonisation of the wasted coffee tray leach

Cristina Vilanova[1], Alba Iglesias[1] & Manuel Porcar[1,2]

Microbial communities are ubiquitous in both natural and artificial environments. However, microbial diversity is usually reduced under strong selection pressures, such as those present in habitats rich in recalcitrant or toxic compounds displaying antimicrobial properties. Caffeine is a natural alkaloid present in coffee, tea and soft drinks with well-known antibacterial properties. Here we present the first systematic analysis of coffee machine-associated bacteria. We sampled the coffee waste reservoir of ten different Nespresso machines and conducted a dynamic monitoring of the colonization process in a new machine. Our results reveal the existence of a varied bacterial community in all the machines sampled, and a rapid colonisation process of the coffee leach. The community developed from a pioneering pool of enterobacteria and other opportunistic taxa to a mature but still highly variable microbiome rich in coffee-adapted bacteria. The bacterial communities described here, for the first time, are potential drivers of biotechnologically relevant processes including decaffeination and bioremediation.

Sample n°.	Machine model	Use*
1	*Essenza*	Communal (12 uses/day)
2	*Pixie*	Domestic (3 uses/day)
3	*Essenza*	Domestic (2 uses/day)
4	*CityZ* (capsule container)	Communal (10 uses/day)
5	*CityZ* (drip base)	
6	*Essenza*	Communal (20 uses/day)
7	*Inissia*	Domestic (4 uses/day)
8	*Essenza*	Communal (10 uses/day)
9	*Pixie*	Domestic (4 uses/day)
10	*Pixie*	Domestic (3 uses/day)
11–30	*Inissia* (Krups)	This work (3 uses/day)

Table 1. Nespresso machines sampled in this work. *Frequency of use indicated in parenthesis as the average number of uses per day, as stated by the users of each machine.

The coffee-machine bacteriome

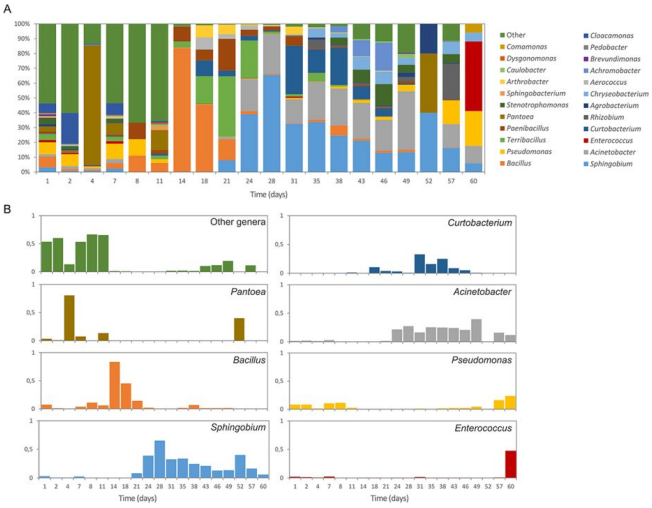


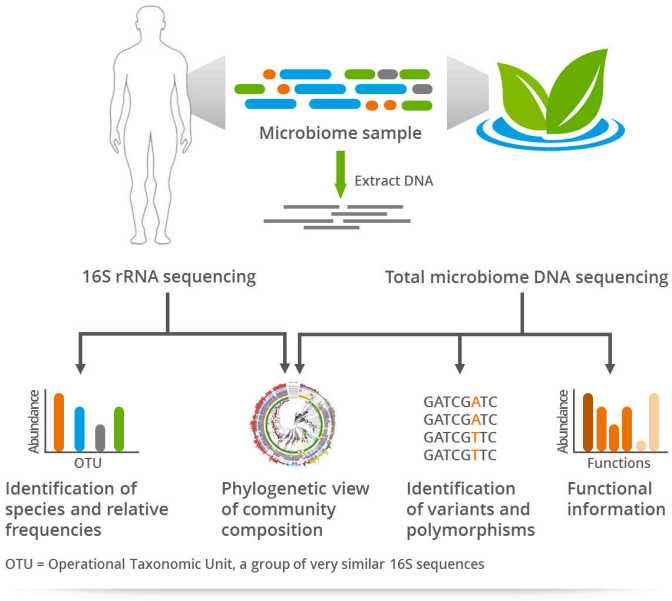
Figure 2. Bacterial colonisation in a brand new Nespresso Krups Inissia machine. (A) Bacterial profile in the drip tray during the two months of operation according to 16S rRNA gene monitoring. (B) Ecological succession of the main taxa during the experiment, represented as the variation of their relative frequencies.

Group Activity

Make groups of four. As a group, take ten minutes to write down answers to each of the following questions:

1. Write down a **definition of *meta*genomics**
2. How would you **count the number of species** in 1L of sea water from Somorrostro beach?
3. How would you determine **how many viable individual bacterial cells** are in 1L of sea water from Somorrostro beach? What are some benefits and problems of your method? (Why is it good? What might it miss?)
4. How would you determine **how many unique bacterial species** are there in 1L of sea water from Somorrostro beach?
   - Without sequencing? - With sequencing?
   hints for sequencing: different species have different alleles (versions) of each gene. some genes, but not all genes, are unique to each species.

Sequencing to identify species



16S ribosomal RNA

16S ribosomal RNA (or 16S rRNA) is the component of the 30S small subunit of a prokaryotic ribosome that binds to the Shine-Dalgarno sequence. The genes coding for it are referred to as **16S rRNA gene** and are used in reconstructing phylogenies, due to the slow rates of evolution of this region of the gene.[2] Carl Woese and George E. Fox were two of the people who pioneered the use of 16S rRNA in phylogenetics.[3]

Multiple sequences of the 16S rRNA gene can exist within a single bacterium.[4]

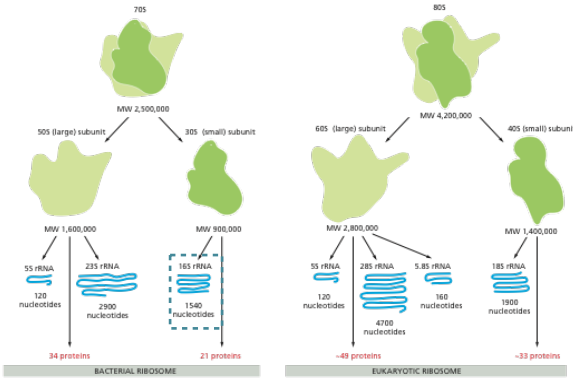


Figure 6–61 **A comparison of bacterial and eukaryotic ribosomes.** Despite differences in the number and size of their rRNA and protein components, both bacterial and eukaryotic ribosomes have nearly the same structure and they function similarly. Although the 16S and 28S rRNAs of the eukaryotic ribosome contain many nucleotides not present in their bacterial counterparts, these nucleotides are present as multiple insertions that form extra domains and leave the basic structure of the rRNA largely unchanged.

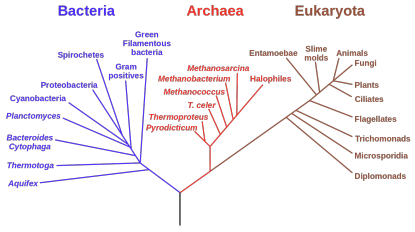
Which phylogenetic domain contains 16S rDNA?

What is the equivalent gene in another domain of life?

Phylogenetic tree of life

3 domains

6-8 kingdoms



Linnaeus 1735	Haeckel 1866	Chatton 1925	Copeland 1938	Whittaker 1969	Woese *et al.* 1977	Woese *et al.* 1990	Cavalier-Smith 1993	Cavalier-Smith 1998	Ruggiero *et al.* 2015
2 kingdoms	3 kingdoms	2 empires	4 kingdoms	5 kingdoms	6 kingdoms	3 domains	8 kingdoms	6 kingdoms	7 kingdoms
(not treated)	Protista	Prokaryota	Monera	Monera	Eubacteria	Bacteria	Eubacteria	Bacteria	Bacteria
					Archaebacteria	Archaea	Archaebacteria		Archaea
		Eukaryota	Protista	Protista	Protista	Eucarya	Archezoa	Protozoa	Protozoa
							Protozoa		
							Chromista	Chromista	Chromista
Vegetabilia	Plantae		Plantae	Plantae	Plantae		Plantae	Plantae	Plantae
				Fungi	Fungi		Fungi	Fungi	Fungi
Animalia	Animalia		Animalia	Animalia	Animalia		Animalia	Animalia	Animalia

16S ribosomal RNA

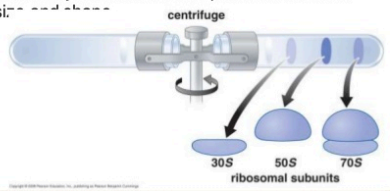
What does the *S* in 16S stand for?
What are the units of this measure? nucleotides? seconds? amino acids?

16S ribosomal RNA

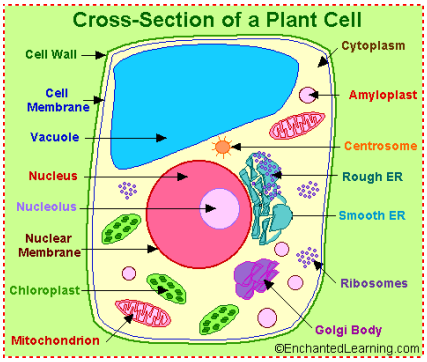
What does the *S* in 16S stand for?
What are the units of this measure? nucleotides? seconds? amino acids?

Svedberg Unit (S)

- The large and small subunit of ribosome are named according to the velocity of sedimentation when subjected to centrifugal force.
- The unit used to measure sedimentation velocity is **Sevdberg (S).**
- The larger value faster is the sedimentation velocity, hence larger the molecule.
- Named after inventor of ultracentrifuge Theodor Sevdberg.
- The sedimentation velocity is a function of a particles molecular weight, volume, si[illegible]



16S ribosomal RNA



The bacterial ribosome has a 16S rRNA, while the eukaryotic ribosome has an 18S rRNA.
Is there a 16S sequence in eukaryotic cells?

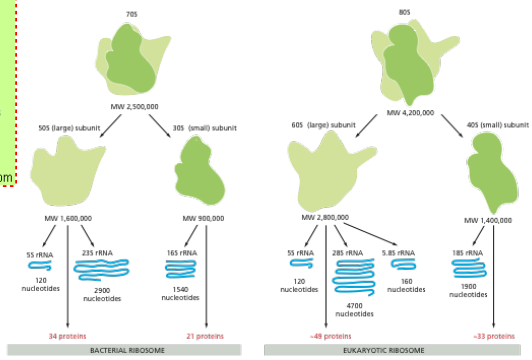


Figure 6-61 **A comparison of bacterial and eukaryotic ribosomes.** Despite differences in the number and size of their rRNA and protein components, both bacterial and eukaryotic ribosomes have nearly the same structure and they function similarly. Although the 16S and 28S rRNAs of the eukaryotic ribosome contain many nucleotides not present in their bacterial counterparts, these nucleotides are present as multiple insertions that form extra domains and leave the basic structure of the rRNA largely unchanged.

16S ribosomal RNA

Is the 16S rDNA gene present in only one copy in all organisms?

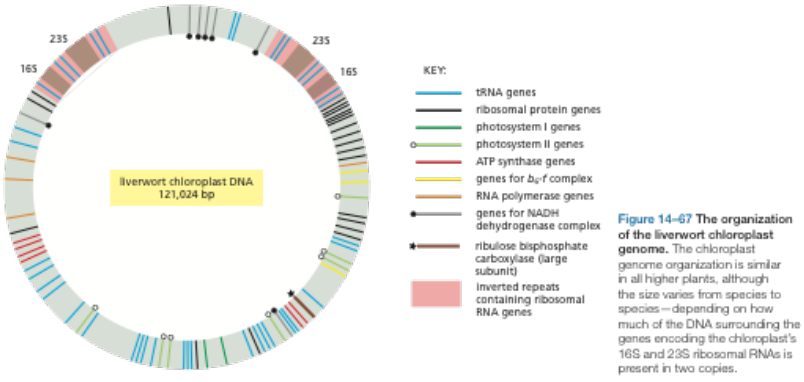
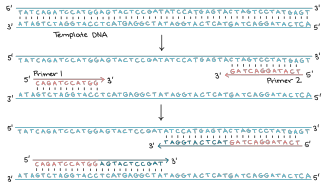
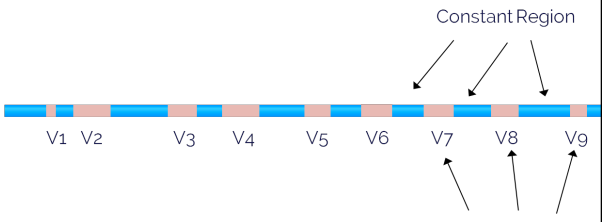


Figure 14–67 **The organization of the liverwort chloroplast genome.** The chloroplast genome organization is similar in all higher plants, although the size varies from species to species—depending on how much of the DNA surrounding the genes encoding the chloroplast's 16S and 23S ribosomal RNAs is present in two copies.

How does 16S sequencing actually work?

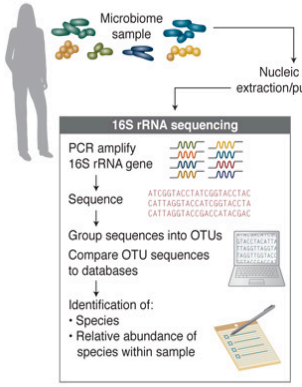
In groups of four, design a PCR & sequencing experiment to determine which species are on the plate.

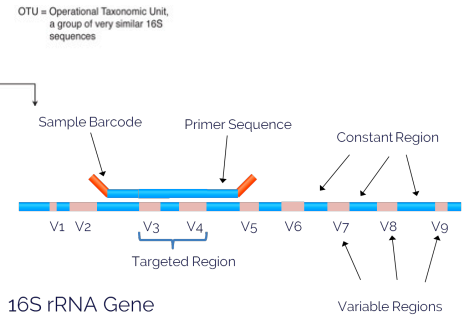




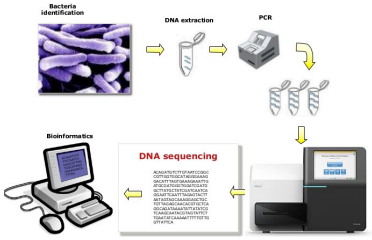
Khan academy

How does 16S sequencing actually work?





Each experiment (eg: each environmental sample) gets a different barcode.
We can then feed multiple samples into the Illumina sequencer.



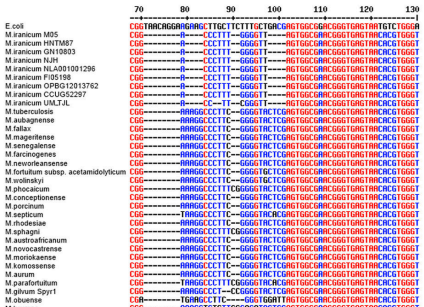
16S rDNA sequences from mycobacterial species

Which species can you uniquely identify from these data? which can you not?

```
                                    70        80        90       100       110       120       130
                                    --+---------+---------+---------+---------+---------+---------|
E.coli                              CGGTAACAGGAAGAAGCTTGCTTCTTTGCTGACGAGTGGCGGACGGGTGAGTAATGTCTGGGA
M.iranicum M05                      CGG--------A-----CCCTTT--GGGGTT----AGTGGCGAACGGGTGAGTAACACGTGGGT
M.iranicum HNTM87                   CGG--------A-----CCCTTT--GGGGTT----AGTGGCGAACGGGTGAGTAACACGTGGGT
M.iranicum GN10803                  CGG--------A-----CCCTTT--GGGGTT----AGTGGCGAACGGGTGAGTAACACGTGGGT
M.iranicum NJH                      CGG--------A-----CCCTTT--GGGGTT----AGTGGCGAACGGGTGAGTAACACGTGGGT
M.iranicum NLA001001296             CGG--------A-----CCCTTT--GGGGTT----AGTGGCGAACGGGTGAGTAACACGTGGGT
M.iranicum FI05198                  CGG--------A-----CCCTTT--GGGGTT----AGTGGCGAACGGGTGAGTAACACGTGGGT
M.iranicum OPBG12013762             CGG--------A-----CCCTTT--GGGGTT----AGTGGCGAACGGGTGAGTAACACGTGGGT
M.iranicum CCUG52297               CGG--------A-----CCCTTT--GGGGTT----AGTGGCGAACGGGTGAGTAACACGTGGGT
M.iranicum UM_TJL                   CGG--------A-----CC--TT--CGGGTT----AGTGGCGAACGGGTGAGTAACACGTGGGT
M.tuberculosis                      CGG--------AAAGGCCCTTC--GGGGTACTCGAGTGGCGAACGGGTGAGTAACACGTGGGT
M.aubagnense                        CGG--------AAAGGCCCTTC--GGGGTACTCGAGTGGCGAACGGGTGAGTAACACGTGGGT
M.fallax                            CGG--------AAAGGCCCTTC--GGGGTACTCGAGTGGCGAACGGGTGAGTAACACGTGGGT
M.mageritense                       CGG--------AAAGGCCCTTC--GGGGTACTCGAGTGGCGAACGGGTGAGTAACACGTGGGT
M.senegalense                       CGG--------AAAGGCCCTTC--GGGGTACTCGAGTGGCGAACGGGTGAGTAACACGTGGGT
M.farcinogenes                      CGG--------AAAGGCCCTTC--GGGGTACTCGAGTGGCGAACGGGTGAGTAACACGTGGGT
M.neworleansense                    CGG--------AAAGGCCCTTC--GGGGTACTCGAGTGGCGAACGGGTGAGTAACACGTGGGT
M.fortuitum subsp. acetamidolyticum CGG--------AAAGGCCCTTC--GGGGTGCTCGAGTGGCGAACGGGTGAGTAACACGTGGGT
M.wolinskyi                         CGG--------AAAGGCCCTTC--GGGGTGCTCGAGTGGCGAACGGGTGAGTAACACGTGGGT
M.phocaicum                         CGG--------AAAGGCCCTTTCGGGGGTACTCGAGTGGCGAACGGGTGAGTAACACGTGGGT
M.conceptionense                    CGG--------AAAGGCCCTTC--GGGGTACTCGAGTGGCGAACGGGTGAGTAACACGTGGGT
M.porcinum                          CGG--------AAAGGCCCTTC--GGGGTACTCGAGTGGCGAACGGGTGAGTAACACGTGGGT
M.septicum                          CGG--------TAAGGCCCTTC--GGGGTACACGAGTGGCGAACGGGTGAGTAACACGTGGGT
M.rhodesiae                         CGG--------AAAGGCCCTTC--GGGGTACTCGAGTGGCGAACGGGTGAGTAACACGTGGGT
M.sphagni                           CGG--------AAAGGCCCTTTCGGGGGTACTCGAGTGGCGAACGGGTGAGTAACACGTGGGT
M.austroafricanum                   CGG--------AAAGGCCCTTC--GGGGTACTCGAGTGGCGAACGGGTGAGTAACACGTGGGT
M.novocastrense                     CGG--------AAAGGCCCTTC--GGGGTACTCGAGTGGCGAACGGGTGAGTAACACGTGGGT
M.moriokaense                       CGG--------AAAGGCCCTTC--GGGGTACTCGAGTGGCGAACGGGTGAGTAACACGTGGGT
M.komossense                        CGG--------AAAGGCCCTTC--GGGGTACTCGAGTGGCGAACGGGTGAGTAACACGTGGGT
M.aurum                             CGG--------AAAGGCCCTTC--GGGGTACTCGAGTGGCGAACGGGTGAGTAACACGTGGGT
M.parafortuitum                     CGG--------TAAGGCCCTTTCGGGGGTACACGAGTGGCGAACGGGTGAGTAACACGTGGGT
M.gilvum Spyr1                      CGG--------AAAGGCCCT--CCGGGGTACTCGAGTGGCGAACGGGTGAGTAACACGTGGGT
M.obuense                           CGA--------TGAAGCCTTC-----GGGTGGATTAGTGGCGAACGGGTGAGTAACACGTGGGT
M.brumae                            CGG--------AAAGGTCTCTTCGGAGATACTCGAGTGGCGAACGGGTGAGTAACACGTGGGT
```

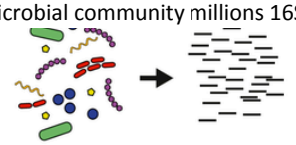
16S rDNA sequences from mycobacterial species

We're not there yet.
Now we have millions of sequencing reads from the 16S genes for all our bacteria
What steps are missing to convert our 16S reads to population frequencies?

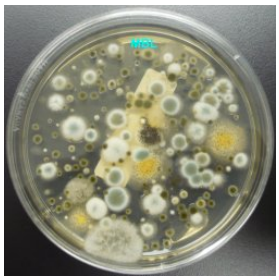


+

microbial community millions 16S reads



?



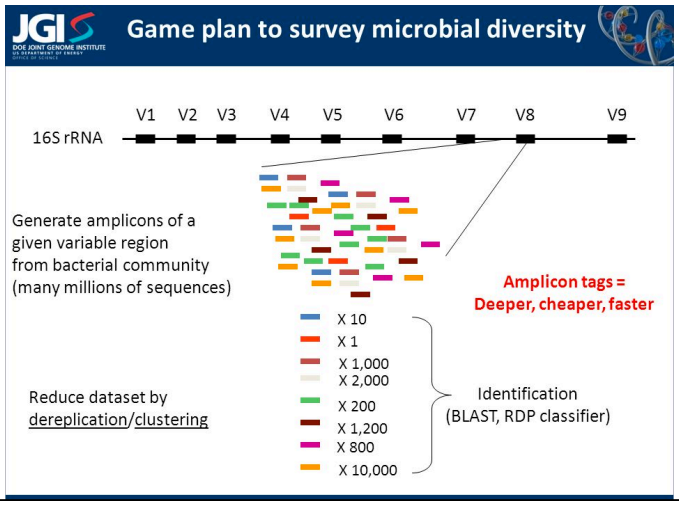
20% of the population

5% of the population

50% of the population

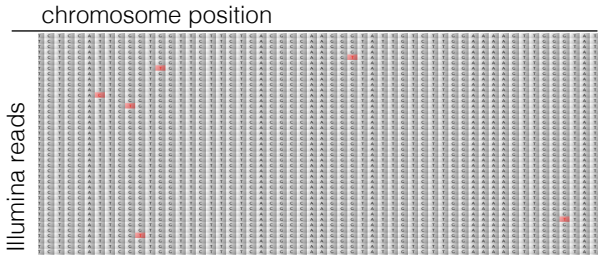
16S rDNA sequences from mycobacterial species

Two options:
(1) map (align) Illumina reads to known 16S sequences from known species
(2) cluster similar sequence reads, and assume that each cluster represents one species (closely related group of organisms)

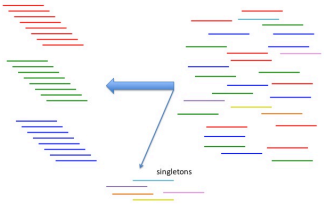


Problem: Illumina sequencing is not error free

Why are sequencing (and PCR amplification) errors a problem when counting species using 16S read counts? How do we solve this problem?
hint: will individual errors be more or less common than individual species?



Clustering (and picking OTUs)



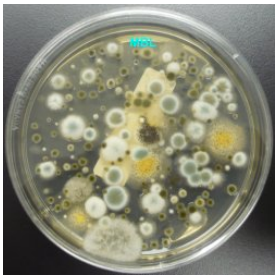
What does an error rate of 10^{-3} mean?

error rates

DNA polymerase	10^{-8}
RNA polymerase	10^{-5}
reverse transcriptase	10^{-5}
Illumina:	10^{-3}

Moving to whole genomes

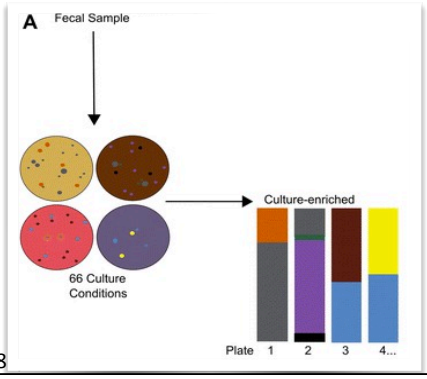
The obvious method:



Isolate unique species

Do whole-genome sequencing of each species

hint:



What is a problem with this approach?

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC492978

Extracting genomes

Solution: assemble into contigs as best you can, then search databases of sequenced genomes. **But we want to find NEW stuff**.

...you only see what is in the database

Extracting genomes

Binning

Genomes from different species have different statistical properties

Genomic signatures:
- GC / Codon usage
- Tetranucleotide frequency + statistical method

PhD student

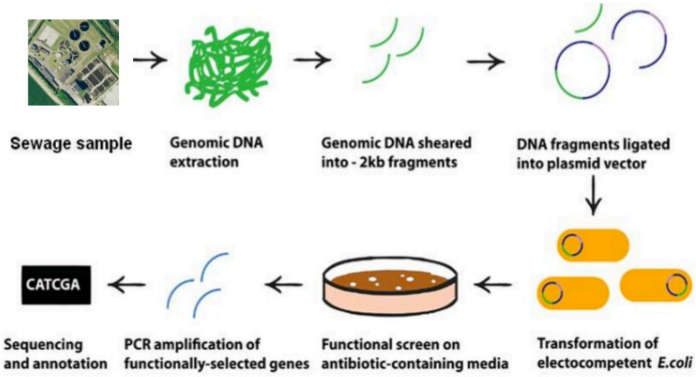
Complex sample

"Binning"

Functional metagenomics

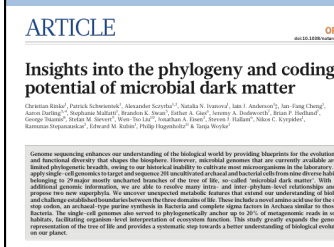
Maybe you don't care about the complete genome?
Just about the functions encoded in genes.

search for novel antibiotics:



M. Sommer, DTU, Denmark (in prep)

Activity 1

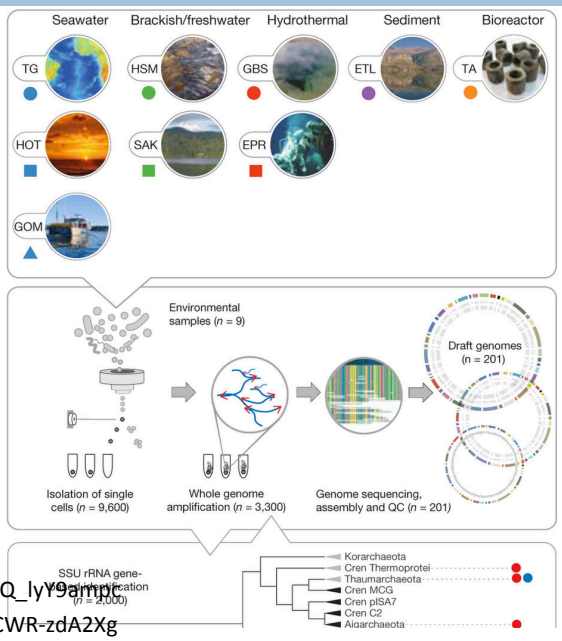


ARTICLE OPEN

Insights into the phylogeny and coding potential of microbial dark matter

Christian Rinke[1], Patrick Schwientek[1], Alexander Sczyrba[1,2], Natalia N. Ivanova[1], Iain J. Anderson[1], Jan-Fang Cheng[1], Aaron Darling[3,4], Stephanie Malfatti[1], Brandon K. Swan[5], Esther A. Gies[6], Jeremy A. Dodsworth[7], Brian P. Hedlund[7], George Tsiamis[8], Stefan M. Sievert[5], Wen-Tso Liu[9], Jonathan A. Eisen[9], Steven J. Hallam[6], Nikos C. Kyrpides[1], Ramunas Stepanauskas[5], Edward M. Rubin[1], Philip Hugenholtz[10] & Tanja Woyke[1]

Genome sequencing enhances our understanding of the biological world by providing blueprints for the evolutionary and functional diversity that shapes the biosphere. However, microbial genomes that are currently available are of limited phylogenetic breadth, owing to our historical inability to cultivate most microorganisms in the laboratory. We apply single-cell genomics to target and sequence 201 uncultivated archaeal and bacterial cells from nine diverse habitats belonging to 29 major mostly uncharted branches of the tree of life, so-called 'microbial dark matter'. With this additional genomic information, we are able to resolve many intra- and inter-phylum-level relationships and to propose two new superphyla. We uncover unexpected metabolic features that extend our understanding of biology and challenge established boundaries between the three domains of life. These include a novel amino acid use for the opal stop codon, an archaeal-type purine synthesis in Bacteria and complete sigma factors in Archaea similar to those in Bacteria. The single-cell genomes also served to phylogenetically anchor up to 20% of metagenomic reads in some habitats, facilitating organism-level interpretation of ecosystem function. This study greatly expands the genomic representation of the tree of life and provides a systematic step towards a better understanding of biological evolution on our planet.



What problem in the culture, isolate, sequence method does single-cell whole-genome sequencing solve? Why?

Why did they assembly only 201 unique genomes from 3300 whole-genome sequences?

https://www.youtube.com/watch?v=9Q_lyY9amPc

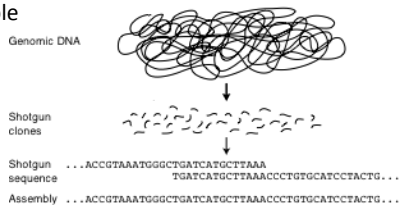
https://www.youtube.com/watch?v=fCWR-zdA2Xg

Activity 2

Couldn't we just :

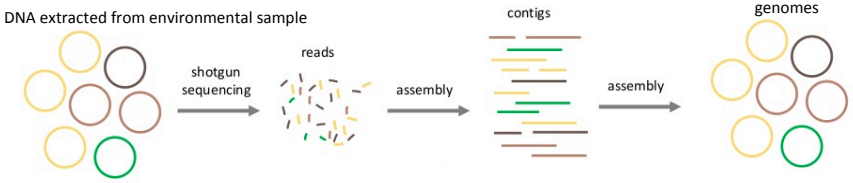
1. isolate all the cells from the environmental sample
2. extract the DNA
3. fragment the DNA and feed it into a sequencer
4. assemble the resulting reads into genomes

Why not?
What problems will this face?



The below pipeline is difficult / impossible. Why?

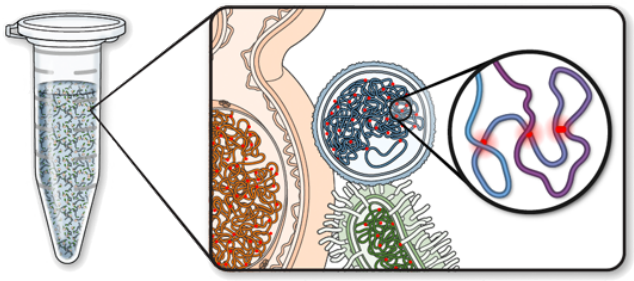
recovering genomes from metagenomic data



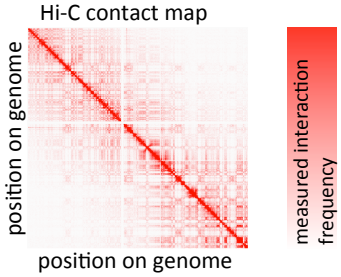
hint: imagine we tried to do this with people to assemble personal genomes. What % of reads would be unique to each individual?

Activity 3

How can we use Hi-C contact maps to solve the assembly problem in metagenomics?



Hi-C contact map



what would a Hi-C contact map look if you mixed two species, and did Hi-C on the mix?