Exploiting sequencing technologies for agriculture

Dr. Orla O’Sullivan
Teagasc Food Research Centre

Orla.osullivan@teagasc.ie
Twitter: @OrlaOS

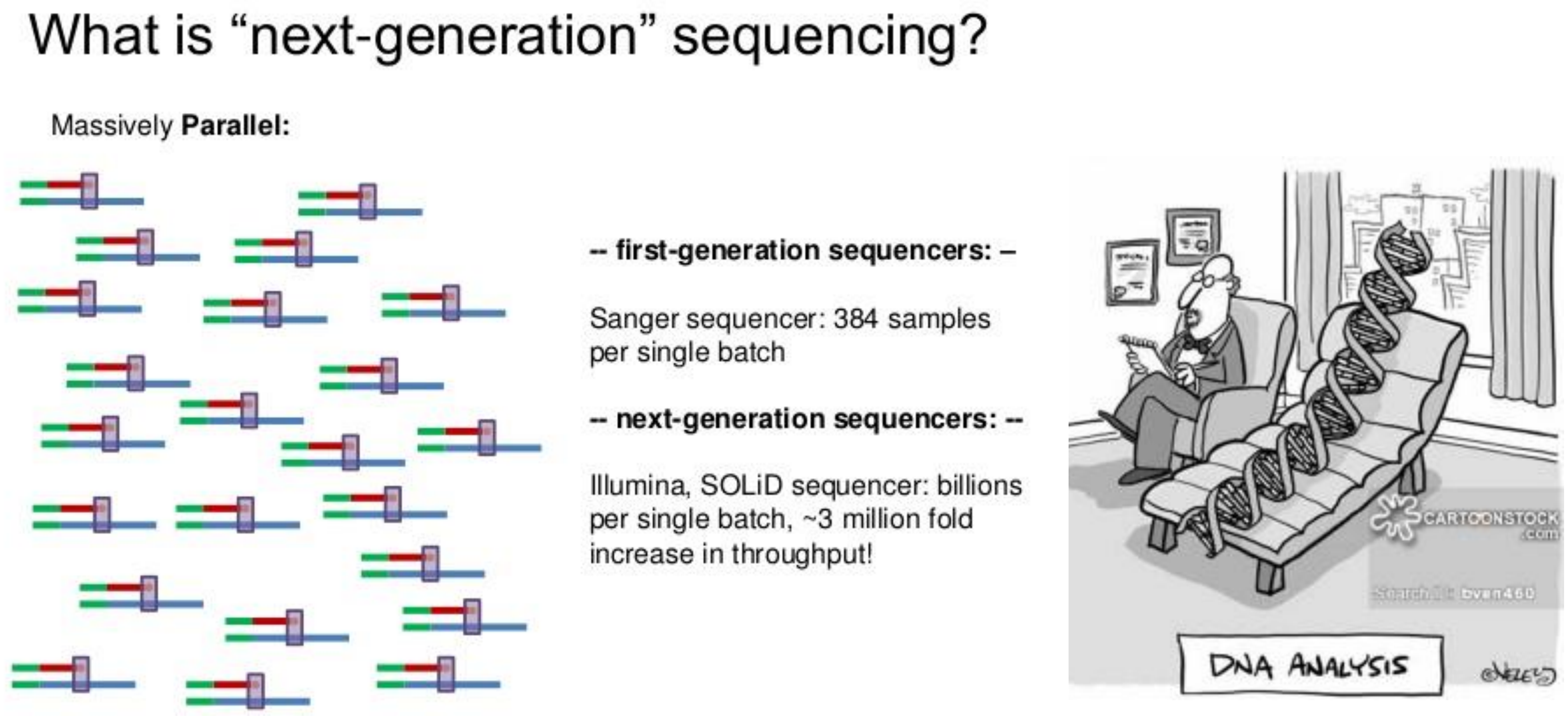
VistaMilk

apc Microbiome Ireland
Interfacing Food & Medicine

Teagasc
Agriculture and Food Development Authority

Science Foundation Ireland sfi For what's next

Next Generation Sequencing



What is “next-generation” sequencing?

Massively **Parallel:**

-- first-generation sequencers: –

Sanger sequencer: 384 samples per single batch

-- next-generation sequencers: --

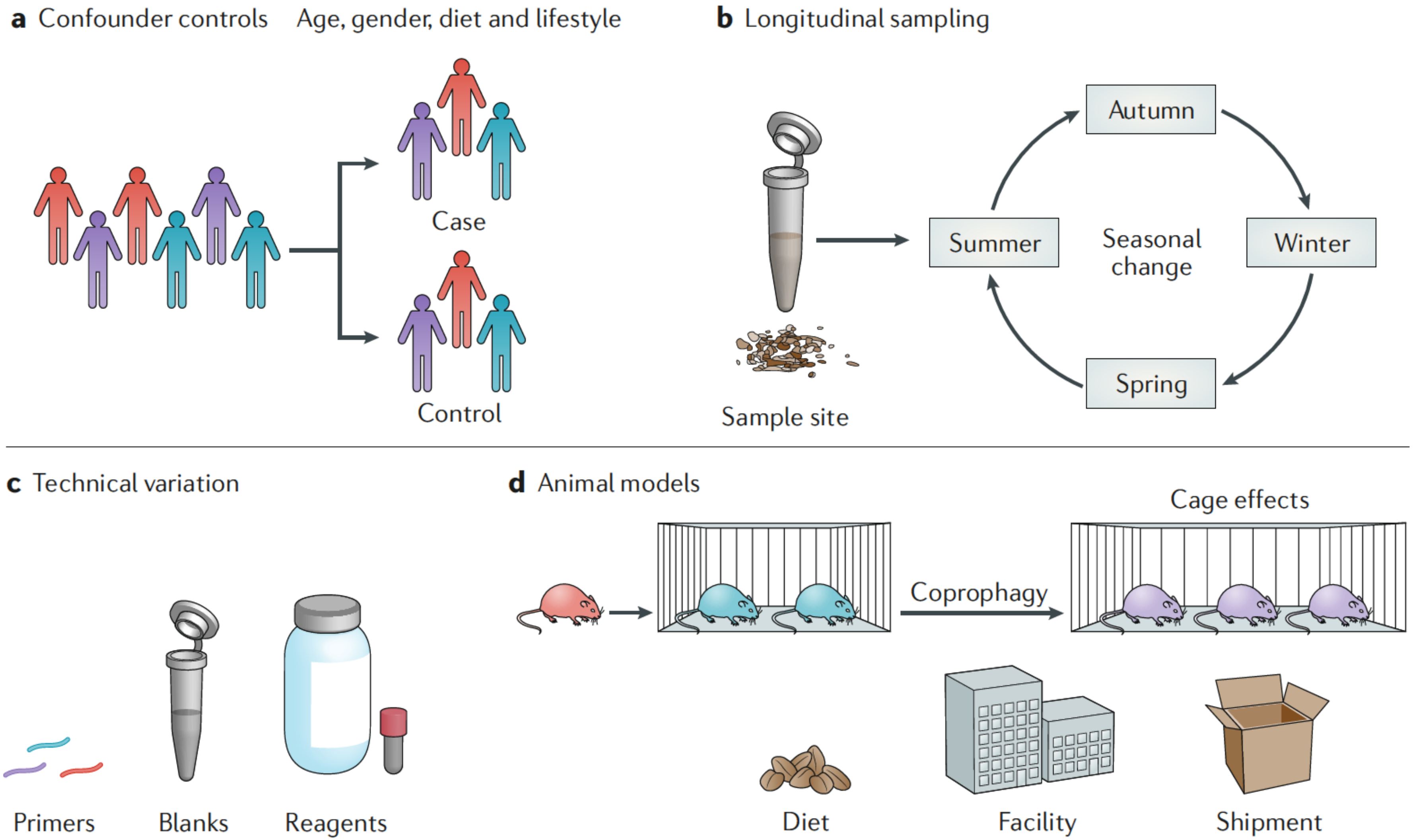
Illumina, SOLiD sequencer: billions per single batch, ~3 million fold increase in throughput!

CARTOONSTOCK .com
Search ID: bven460
DNA ANALYSIS

Plan Plan Plan!!!!

Things to consider in your plan…

- What samples are adequate for biological question(s) and how much sample to adequate profile environment (e.g. soil)
- Is it covered by ethics and who is collecting
- Type of extraction method - Extract DNA/RNA or both
- Adequate aliquots (avoid freeze-thaw cycles)
- Amount of sample available
- Starting amount within optimal range of extraction method
- Extract host versus bacterial DNA/Fungal/viral
- Amount of organic matter/Potential PCR inhibitors in sample
- Randomising order of samples in workflow
- Ensure consistency and record keeping (operator, equipment, lot numbers)

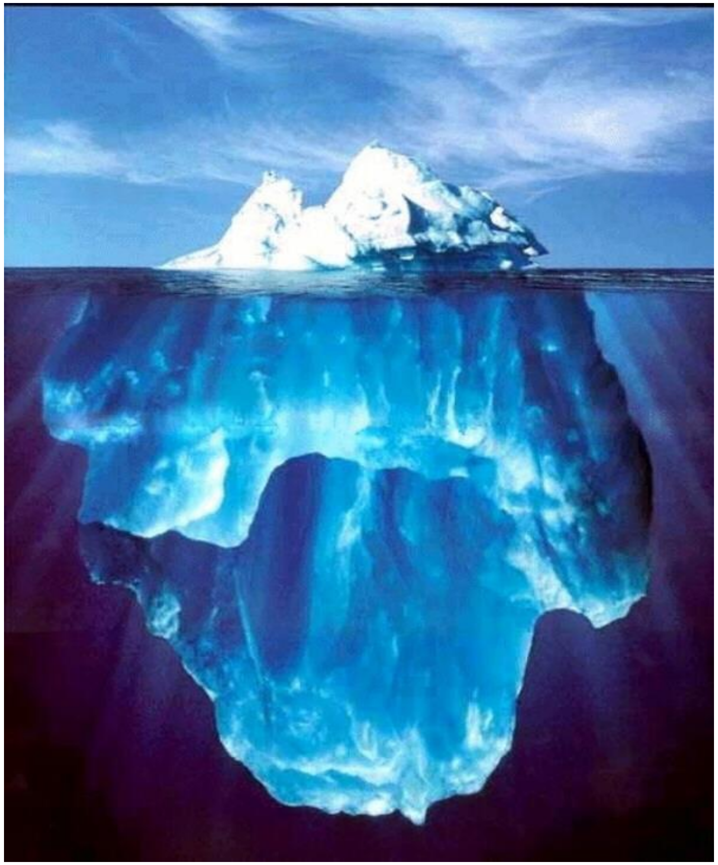


Knight, R.,et al., (2018). "Best practices for analysing microbiomes." Nature Reviews Microbiology.

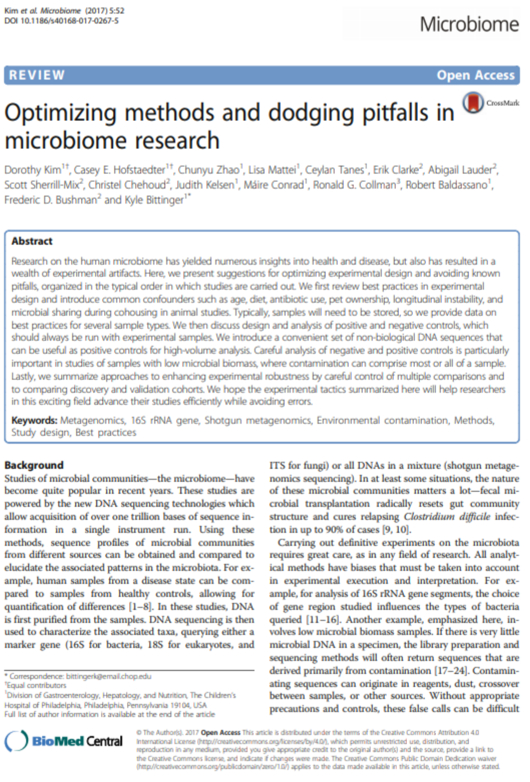
Why Next Generation Sequencing?

THE PLATE COUNT ANOMALY

- Culturable fraction < 30%;



The importance of negative controls in microbiota analysis



Kim et al. Microbiome (2017) 5:52
DOI 10.1186/s40168-017-0267-5

Microbiome

REVIEW Open Access

Optimizing methods and dodging pitfalls in microbiome research

Dorothy Kim[1†], Casey E. Hofstaedter[1†], Chunyu Zhao[1], Lisa Mattei[1], Ceylan Tanes[1], Erik Clarke[2], Abigail Lauder[2], Scott Sherrill-Mix[2], Christel Chehoud[2], Judith Kelsen[1], Máire Conrad[1], Ronald G. Collman[3], Robert Baldassano[1], Frederic D. Bushman[2] and Kyle Bittinger[1*]

Abstract

Research on the human microbiome has yielded numerous insights into health and disease, but also has resulted in a wealth of experimental artifacts. Here, we present suggestions for optimizing experimental design and avoiding known pitfalls, organized in the typical order in which studies are carried out. We first review best practices in experimental design and introduce common confounders such as age, diet, antibiotic use, pet ownership, longitudinal instability, and microbial sharing during cohousing in animal studies. Typically, samples will need to be stored, so we provide data on best practices for several sample types. We then discuss design and analysis of positive and negative controls, which should always be run with experimental samples. We introduce a convenient set of non-biological DNA sequences that can be useful as positive controls for high-volume analysis. Careful analysis of negative and positive controls is particularly important in studies of samples with low microbial biomass, where contamination can comprise most or all of a sample. Lastly, we summarize approaches to enhancing experimental robustness by careful control of multiple comparisons and to comparing discovery and validation cohorts. We hope the experimental tactics summarized here will help researchers in this exciting field advance their studies efficiently while avoiding errors.

Keywords: Metagenomics, 16S rRNA gene, Shotgun metagenomics, Environmental contamination, Methods, Study design, Best practices

Background

Studies of microbial communities—the microbiome—have become quite popular in recent years. These studies are powered by the new DNA sequencing technologies which allow acquisition of over one trillion bases of sequence information in a single instrument run. Using these methods, sequence profiles of microbial communities from different sources can be obtained and compared to elucidate the associated patterns in the microbiota. For example, human samples from a disease state can be compared to samples from healthy controls, allowing for quantification of differences [1–8]. In these studies, DNA is first purified from the samples. DNA sequencing is then used to characterize the associated taxa, querying either a marker gene (16S for bacteria, 18S for eukaryotes, and ITS for fungi) or all DNAs in a mixture (shotgun metagenomics sequencing). In at least some situations, the nature of these microbial communities matters a lot—fecal microbial transplantation radically resets gut community structure and cures relapsing *Clostridium difficile* infection in up to 90% of cases [9, 10].

Carrying out definitive experiments on the microbiota requires great care, as in any field of research. All analytical methods have biases that must be taken into account in experimental execution and interpretation. For example, for analysis of 16S rRNA gene segments, the choice of gene region studied influences the types of bacteria queried [11–16]. Another example, emphasized here, involves low microbial biomass samples. If there is very little microbial DNA in a specimen, the library preparation and sequencing methods will often return sequences that are derived primarily from contamination [17–24]. Contaminating sequences can originate in reagents, dust, crossover between samples, or other sources. Without appropriate precautions and controls, these false calls can be difficult

* Correspondence: bittingerk@email.chop.edu
†Equal contributors
[1]Division of Gastroenterology, Hepatology, and Nutrition, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA
Full list of author information is available at the end of the article

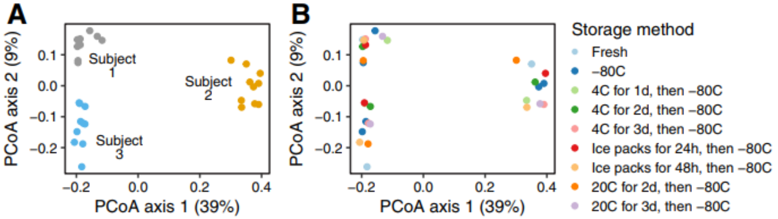


Fig. 2 Effects of sample storage methods on community structure inferred for oral swabs. Oral swab samples were acquired from three human individuals and DNA extracted. DNAs were amplified using 16S rRNA gene primers binding to the V1-V2 region then sequenced using the Illumina platform using our standard procedures [88]. Unweighted Unifrac (C [129].) was used to generate distances between all pairs of samples then results were displayed using Principal Coordinate Analysis (PCoA). **a** Samples from each of the three subjects are color coded (*red*, *blue*, and *green*). **b** Nine storage conditions were compared, indicated by the different *colors*. The key to storage conditions is at the *right*

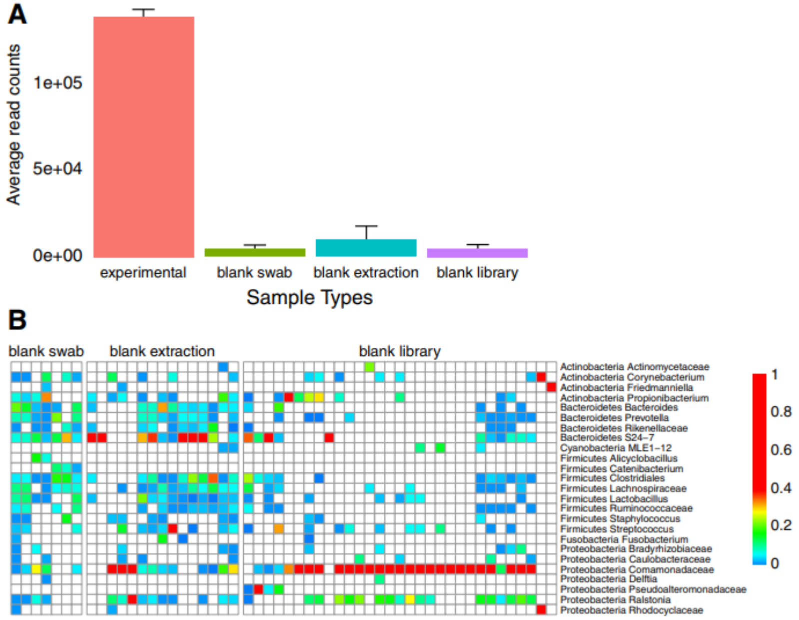


Fig. 4 Analysis of three negative control sample types reveals contaminating taxa. Data for negative controls was acquired using 16S V1-V2 rRNA marker gene sequencing analyzed on the Illumina MiSeq platform. Data from 11 experiments were pooled. **a** Comparison of average read counts. Experimental samples had an average read count of 137,243 and negative control samples had an average read count of 6613. **b** Heat map summary of bacterial lineages present in negative control samples. Different OTUs are present in DNA-extraction controls ("blank extraction" and "blank swab") and library preparation controls ("library blank") collected over multiple sequencing runs

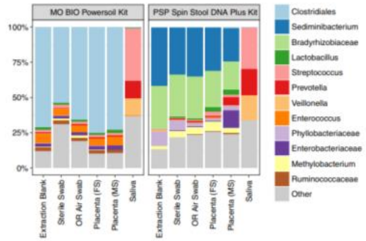


Fig. 3 Wrestling with kit contamination—similar bacterial composition in placental samples and negative controls. Relative abundances of bacterial lineages were inferred from 16S V1-V2 rRNA marker gene sequence information [22]. Samples studied included negative controls, fetal side (FS) placental swabs, maternal side (MS) placental swabs, saliva, and vaginal swabs. Replicates of each sample were extracted using two different kits—the kit type is indicated above each panel. Operating room (OR) air swabs are swabs that were waved in the air at the time of sample collection to be used as negative controls. Saliva samples, which are high in microbial biomass, showed similar compositions for each of the two extractions; placental samples resemble the kit-specific negative controls

The effects of Storage

- Degradation also occurred when frozen samples were defrosted for short periods such as 1 h before nucleic acid extraction.[3]
- DNA and RNA fragment at room temperature for more than 24 hours. [2]

- Immediate freezing at -20 ° C or preservation in RNAlater at room temperature results in similar species compositions. [1]

- Bahl et al, 2011 showed no consistent differences in DNA yield between fresh and frozen samples using 3 different extraction protocols, however they also observed differences in the community composition of frozen samples.
- Maukonen et al, 2011 also reported that the DNA extraction did not affect the diversity, composition, or quantity of Bacteroides spp., but found that after a week's storage at 20 ° C, the numbers of Bacteroides spp. were decreased.

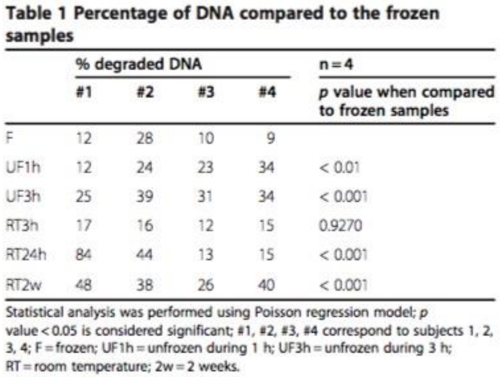


Table 1 Percentage of DNA compared to the frozen samples

	% degraded DNA				n = 4
	#1	#2	#3	#4	*p* value when compared to frozen samples
F	12	28	10	9	
UF1h	12	24	23	34	< 0.01
UF3h	25	39	31	34	< 0.001
RT3h	17	16	12	15	0.9270
RT24h	84	44	13	15	< 0.001
RT2w	48	38	26	40	< 0.001

Statistical analysis was performed using Poisson regression model; *p* value < 0.05 is considered significant; #1, #2, #3, #4 correspond to subjects 1, 2, 3, 4; F = frozen; UF1h = unfrozen during 1 h; UF3h = unfrozen during 3 h; RT = room temperature; 2w = 2 weeks.

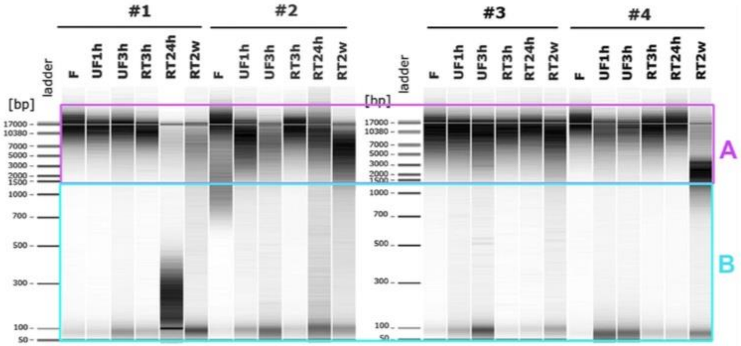


Figure 1 Fragmentation analysis of genomic DNA. Microcapillary electrophoresis patterns of genomic DNA extracted from fecal samples collected by 4 individuals (#1, #2, #3, #4) and stored in the following conditions: immediately frozen at −20°C (F); immediately frozen and then unfrozen during 1 h and 3 h (UF1h, UF3h); kept at room temperature during 3 h, 24 h and 2 weeks (RT3h, RT24h, RT2w). The equivalent to 1 mg of fecal material is loaded on each lane. A DNA fragment size (base pair) ladder was loaded in the left most lanes.

[1] **Voigt *et al.*, 2015**. Temporal and technical variability of human gut metagenomes. *Genome Biology.*
[2] **Franzosa *et al.*, 2015**. Relating the metatranscriptome and metagenome of the human gut. *PNAS.*
[3] **Cardona *et al.*, 2012**. Storage conditions of intestinal microbiota matter in metagenomic analysis. *BMC Microbiol.*

Types of Sequencers

Illumina NextSeq

Minion

Illumina MiSeq

PacBio

Ion Proton

Ion PGM

Illumina HiSeq

SoliD

teagasc

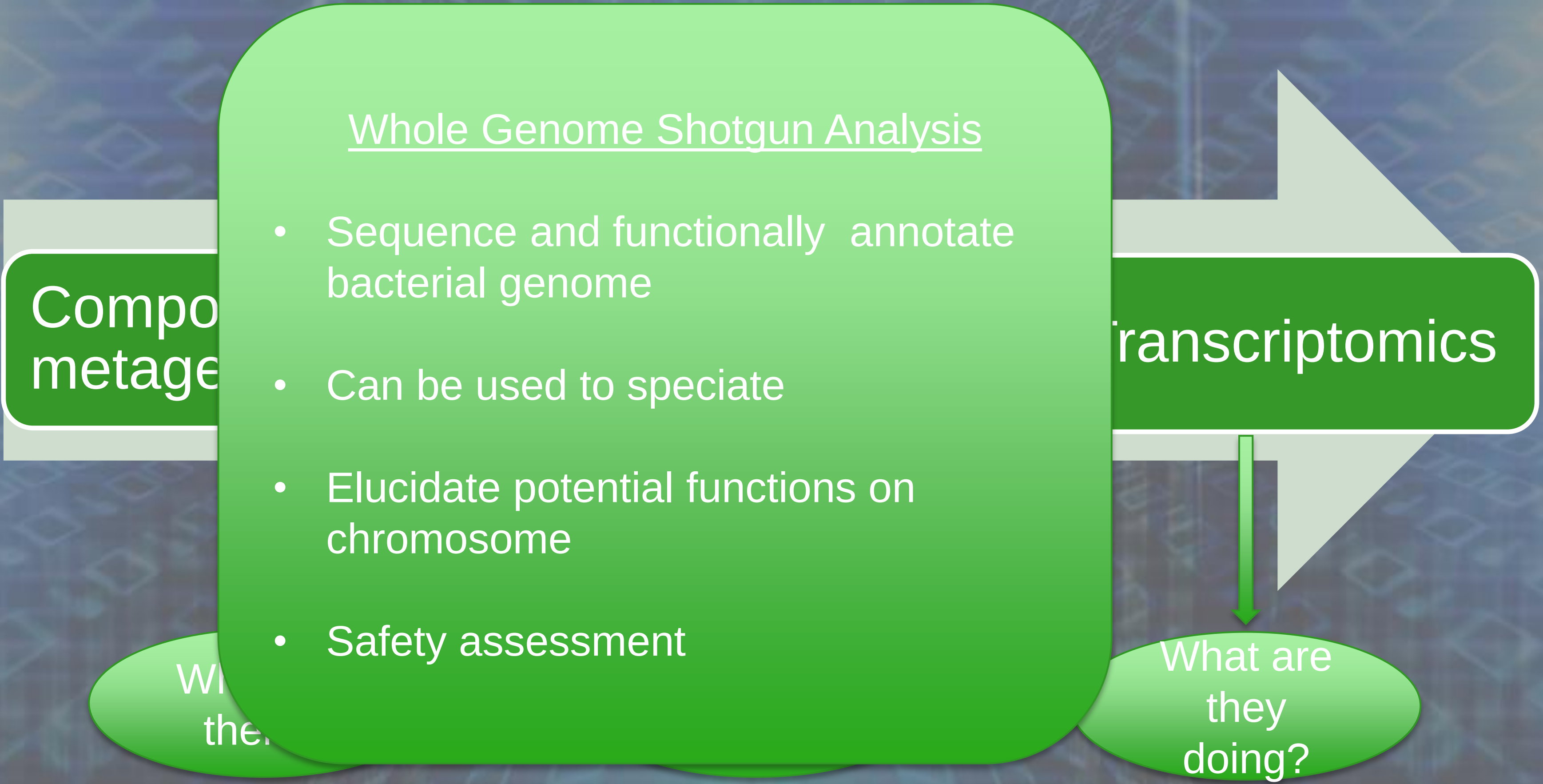
AGRICULTURE AND FOOD DEVELOPMENT AUTHORITY

Which Sequencer to choose?

- Cost
- What question are you asking
- Time
- Availability
- Local expertise – sample prep and downstream analysis

Agriculture and Food Development Authority

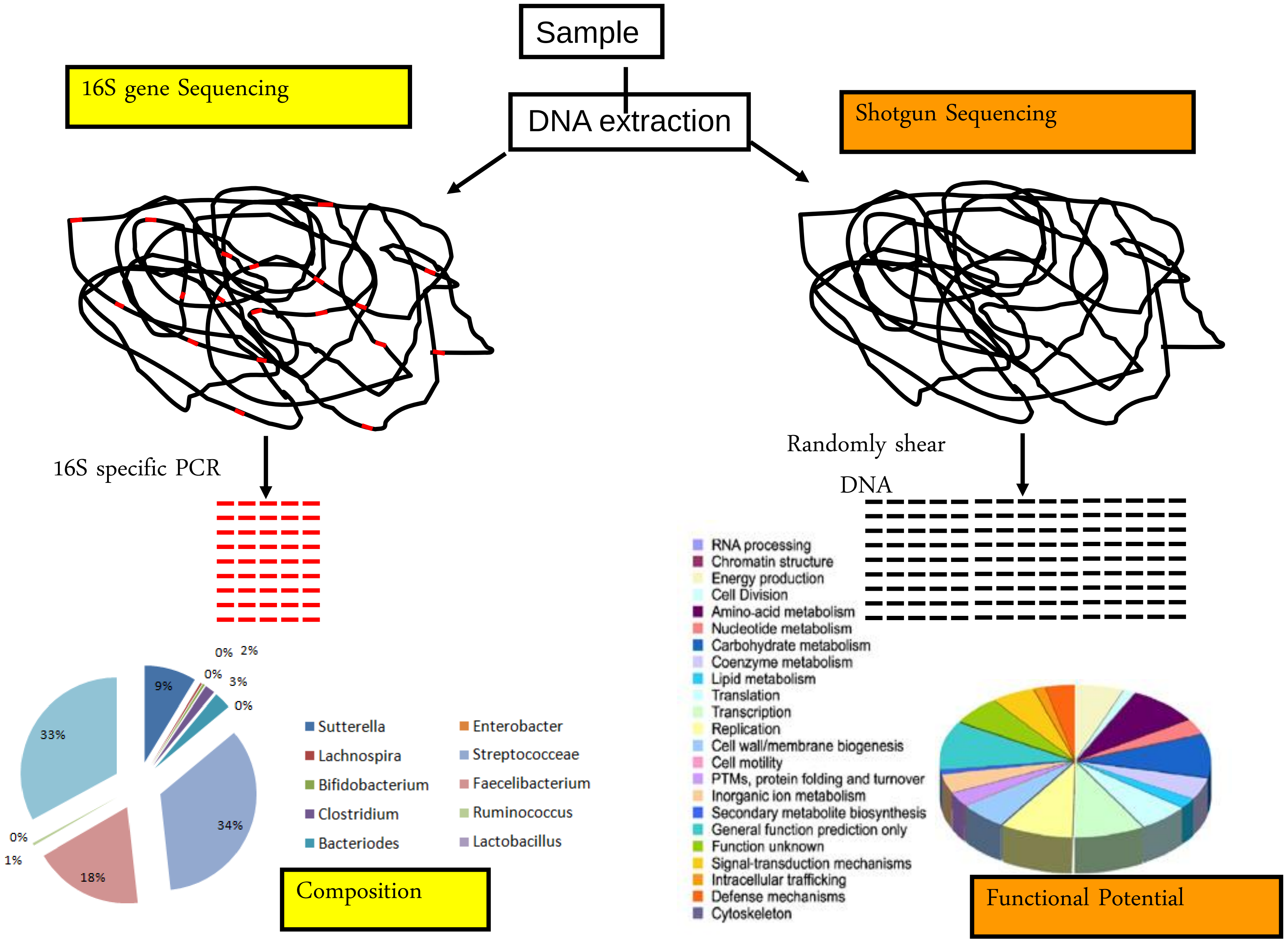
What can we do ?



Whole Genome Shotgun Analysis

- Sequence and functionally annotate bacterial genome
- Can be used to speciate
- Elucidate potential functions on chromosome
- Safety assessment

NGS – METAGENOMICS



Sample

DNA extraction

16S gene Sequencing

Shotgun Sequencing

16S specific PCR

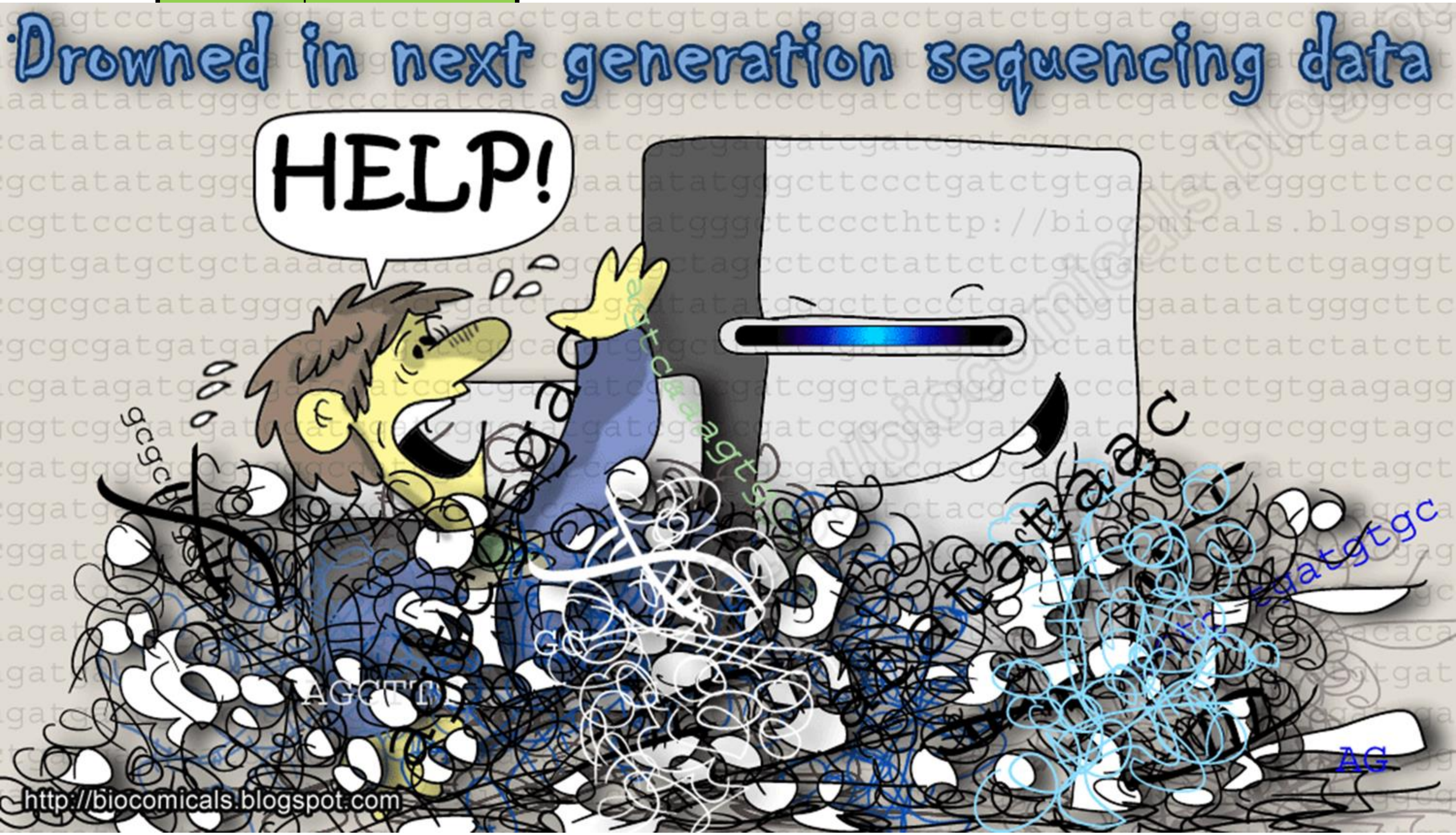
Randomly shear DNA

Composition

Functional Potential

NGS – METATRANSCRIPTOMICS

Metatranscriptomics

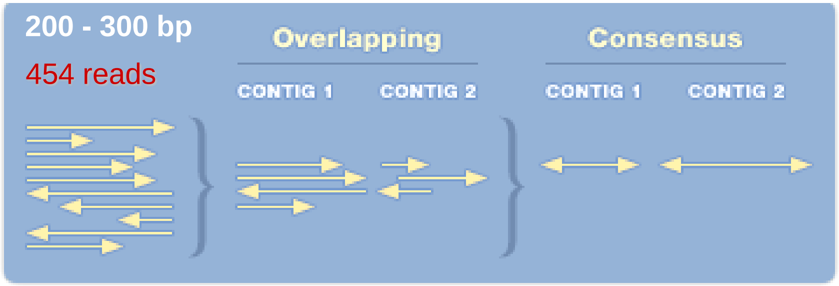

Drowned in next generation sequencing data
HELP!
http://biocomicals.blogspot.com

Gene expression

WHOLE GENOME SEQUENCING

454 example:

Reads ➲ Contigs



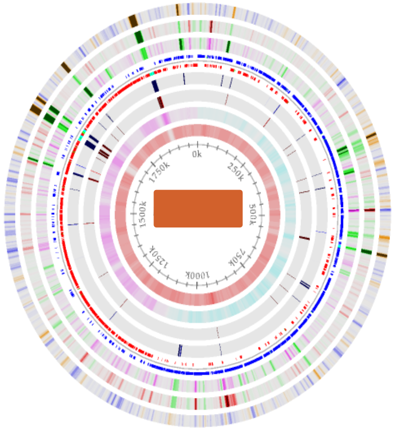
- assembling of reads in to contigs and contigs are ordered in a genome

Sequence quality assessment

Contigs & Scaffold assembly

Gaps of the genome closed by PCR

Functional annotation



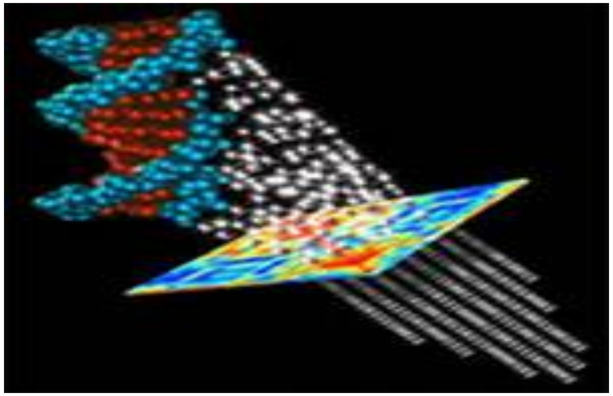
***De novo* sequence**

Teagasc

Agriculture and Food Development Authority

Which type of sequencing?

1) What question are you asking?
2) How much money do you have?
3) Access to computing power
4) Level of bioinformatic support



Agriculture and Food Development Authority

How to use metadata

- Metadata is data about your data

- Making your data avaliable and accessible for others

- Metadata forces you to document better

Planning metadata will allow you to use cohorts from other studies

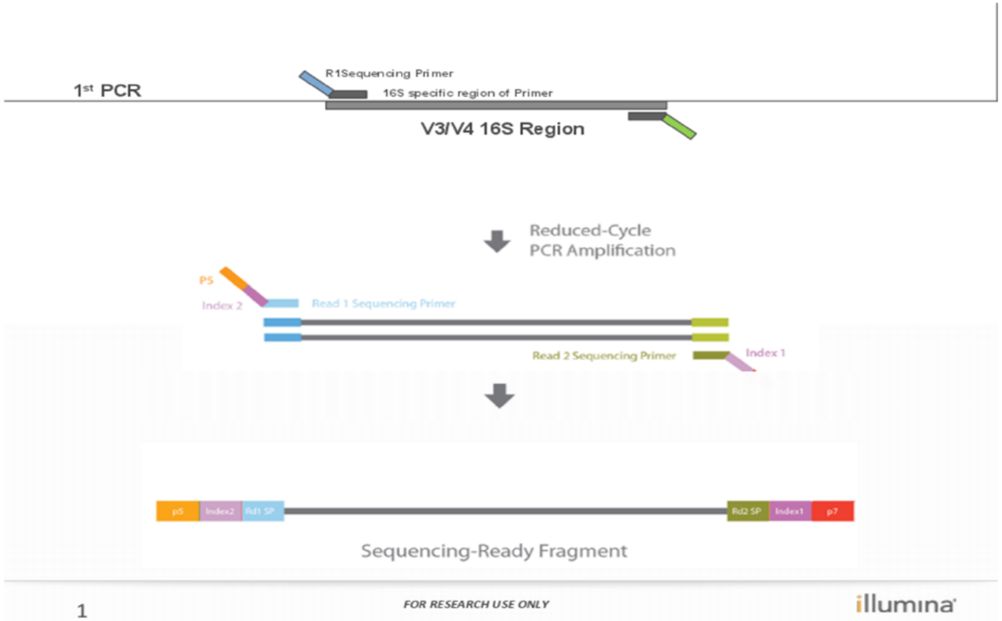
  - Methodology matters!

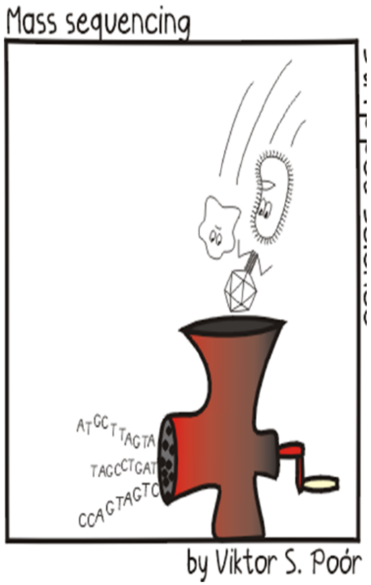
Importance of Metadata

- Powerful if kept consistent and clear
  - **please keep names consistent!**
    - most algorithms are case sensitive
    - for 300 samples, sample 1 should be numbered 001, not 1
    - in excel, coloured cells and ambiguous comments are rarely suited for processing
    - Be careful with spaces, underscores, full stops.
- Bioinformaticians should be involved from initial experimental design
- Good metadata may be difference between a good paper and a great paper

What is amplicon sequencing?

“Sequencing of target genes (are regions of)
obtained by PCR using gene specific primers.”





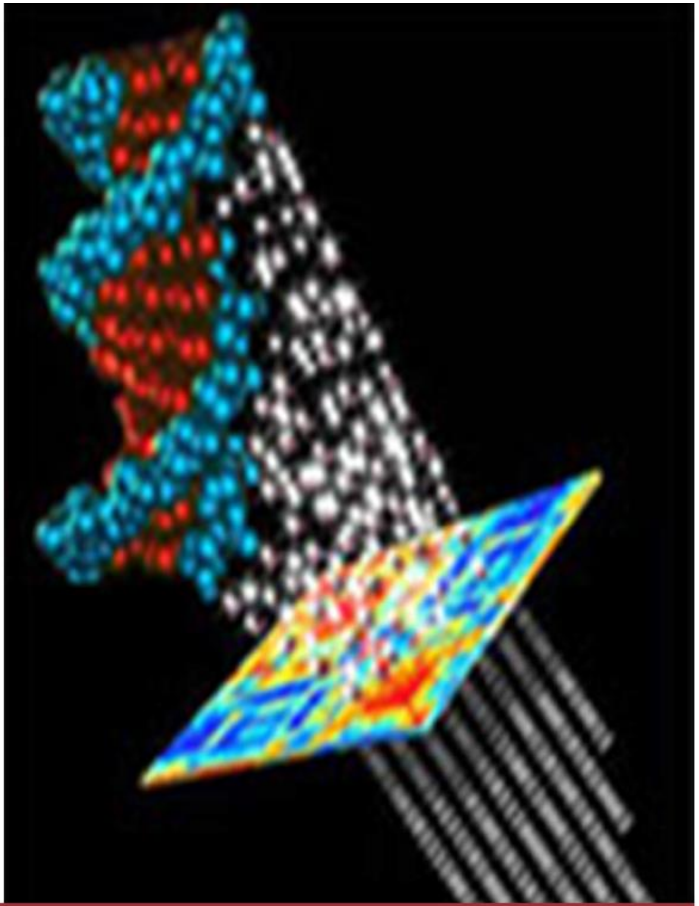
Traits of a marker gene

Traits

- Ubiquitous
- Discriminating
- Slow evolving
- Good database available

Examples

- 16S rRNA gene (bacteria)
- 18S rRNA gene (eukaryotes)
- ITS (fungi)
- Decarboxylases (cheese)
- Target particular species (rpoB; Bifidobacteria)



Amplicon sequencing not suitable for groups such as protists and viruses which are extremely diverse and have little sequence information available

Before you start!

What is your objective ?
e.g. to ascertain what microbial populations are in an environment

What populations do you want to target?
e.g. fungal/bacterial/ archaea ; this will determine primer choice

What is your sample?
e.g. soil, water, cheese, kefir, faeces (mouse, rat, adult, human, fermenter); this will affect extraction protocols

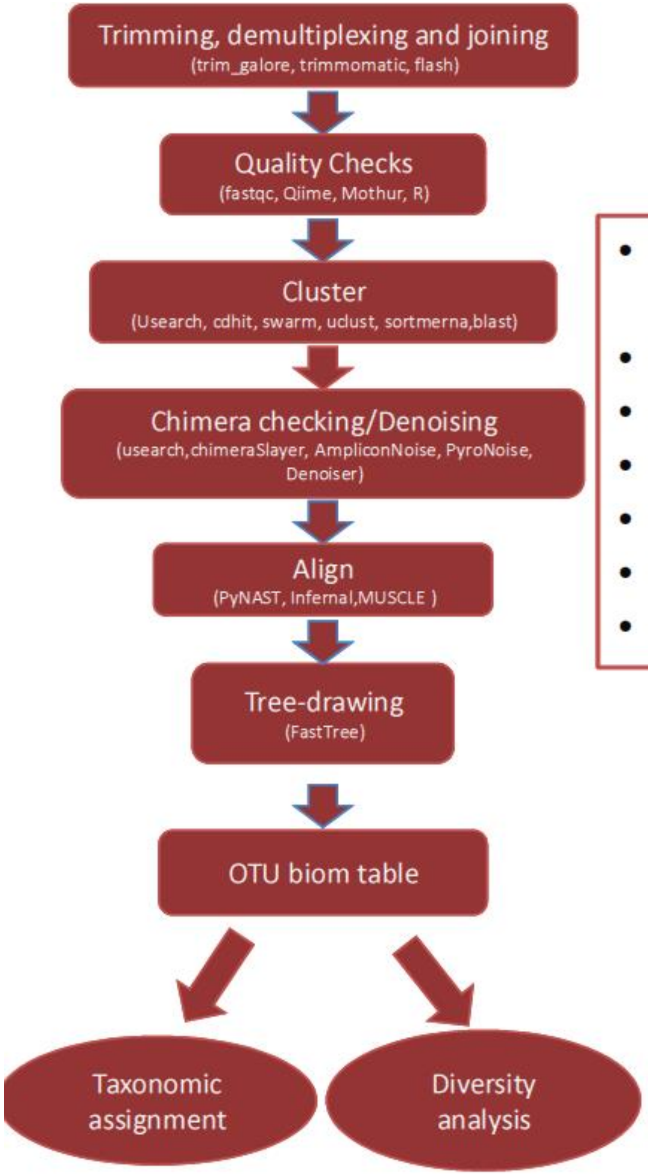
What is in your sample?
Is it a low or high diversity sample; this will affect how much sequencing depth you need
Also number of unique indices available is a limiting factor

Plan, Plan, Plan !!!

Some caveats

- Also sequence dead DNA
- Can miss minor populations
- Use the same platform and primers for each study

Sample Workflow



Trimming, demultiplexing and joining
(trim_galore, trimmomatic, flash)

Quality Checks
(fastqc, Qiime, Mothur, R)

Cluster
(Usearch, cdhit, swarm, uclust, sortmerna,blast)

Chimera checking/Denoising
(usearch,chimeraSlayer, AmpliconNoise, PyroNoise, Denoiser)

Align
(PyNAST, Infernal,MUSCLE)

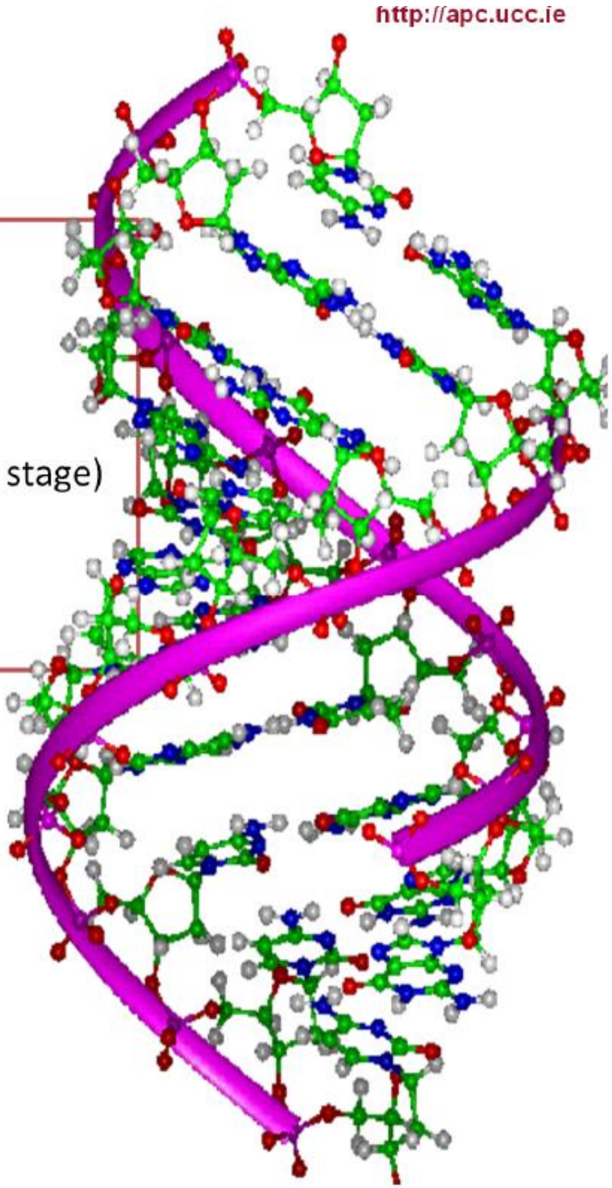
Tree-drawing
(FastTree)

OTU biom table

Taxonomic assignment

Diversity analysis

- **Commonly used complete workflows**
- Qiime (linux)
- Mothur (linux)
- R (numerous packages for each stage)
- Usearch (linux)
- Illumina Basespace
- Greengenes



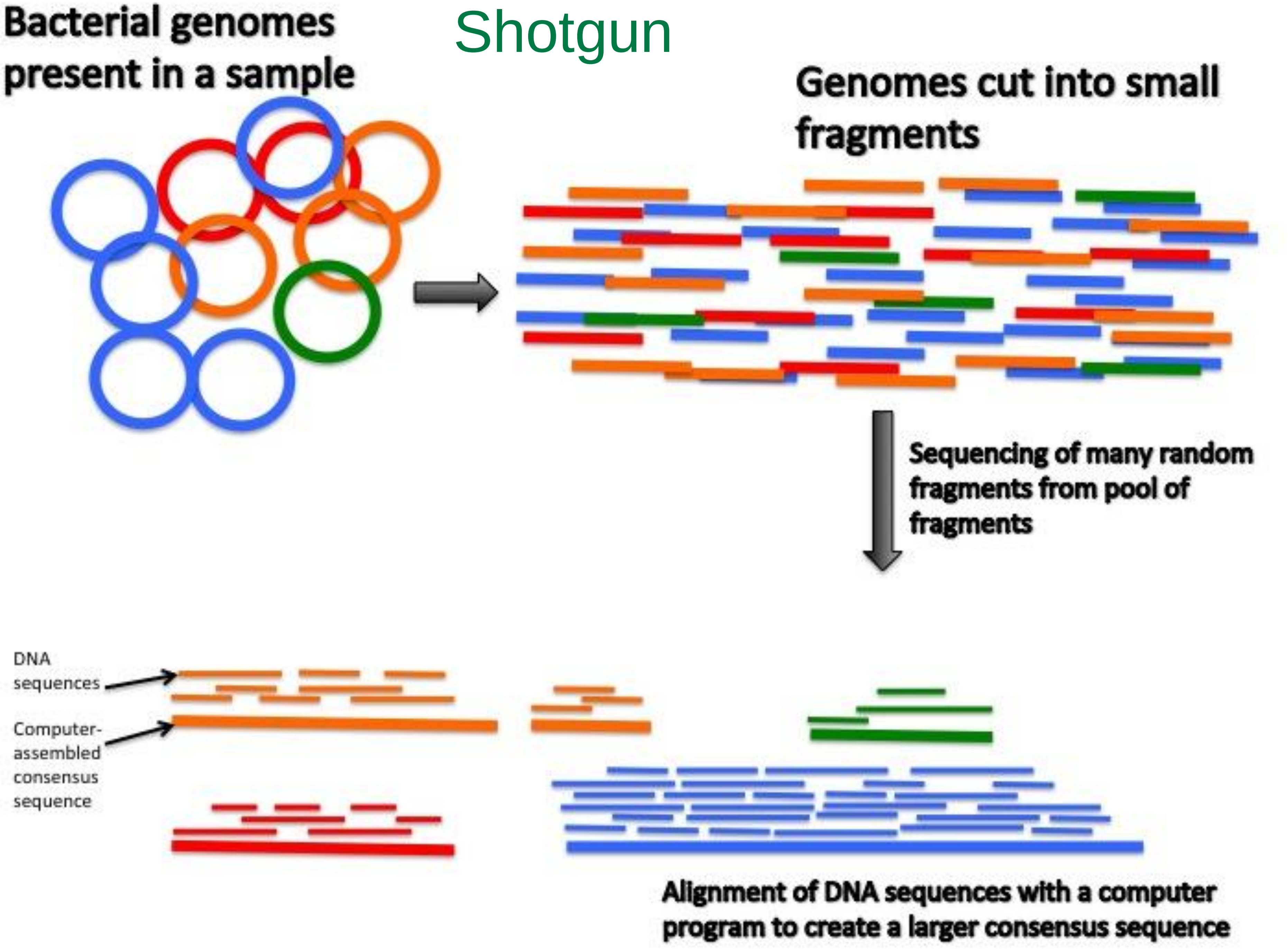
http://apc.ucc.ie

Amplicon or Shotgun Metagenomics?

- Amplicon Cons
  - Different primers will have different detection efficiencies.
  - Sequencing errors may artificially inflate the diversity of the sample.
  - Functional potential can only be predicted indirectly.
  - Often difficult to integrate multiple Amplicon datasets

- Whole shotgun metagenomic Cons
  - Large datasets (Storage)
  - Large datasets (Computation time)
  - Large datasets (Analysis and statistical significance)
  - Cost

Shotgun


Bacterial genomes present in a sample
Genomes cut into small fragments
Sequencing of many random fragments from pool of fragments
DNA sequences
Computer-assembled consensus sequence
Alignment of DNA sequences with a computer program to create a larger consensus sequence

Downstream analysis tools

NB unlike amplicon sequencing there is no standardised methods

- Compositional analysis:
  - Kraken
  - kaiju
  - MetaPhlAn2

- Functional analysis:
  - HUMAnN2
  - SUPER-FOCUS

- Strain-level
  - StrainPhlAn

- Genome reconstruction/assembly
  - MetaBAT
  - Meta-velvet

•Antibiotic resistance –resistome
•Phageome
•Virome

teagasc
Agriculture and Food Development Authority

Intraspecific variation

- Species-level was the best we were able to achieve, until recently

- However, genetic content often varies even within a species

E. coli O157:H7

E. coli Nissle 1917

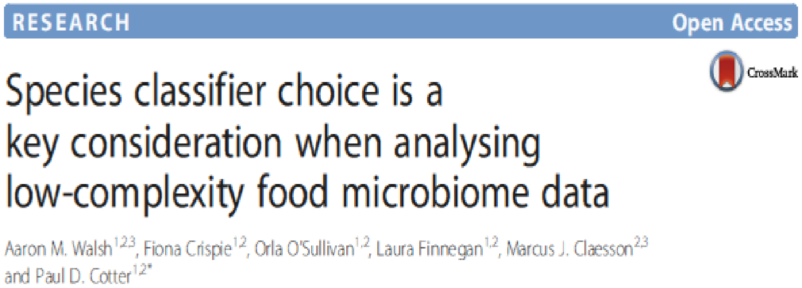
- Ideally, we want to characterise microbes at the strain-level

Strain-level analysis

- Tools for strain-level analysis from shotgun metagenomics:
  - PanPhlAn (10.1038/nmeth.3802)
  - MetaMLST (10.1093/nar/gkw837)
  - StrainPhlAn (10.1101/gr.216242.116)
  - StrainEst (10.1038/s41467-017-02209-5)

Points to ponder

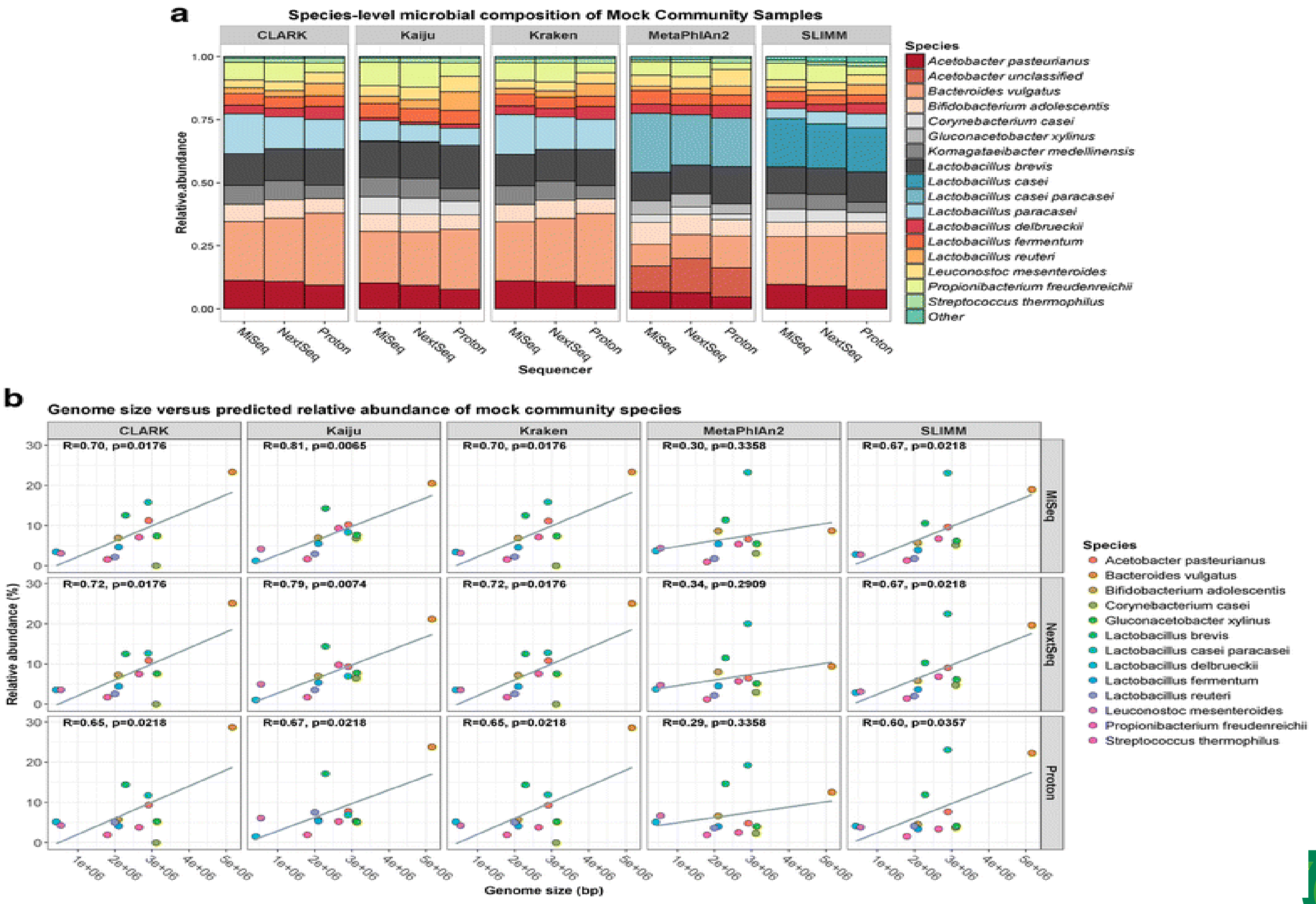
- You will get different answers using different software and databases
- Consistency again is key
- Still get a large number of unassigned
- Assembly may be the key
- Environment is a big factor



RESEARCH Open Access

Species classifier choice is a key consideration when analysing low-complexity food microbiome data

Aaron M. Walsh[1,2,3], Fiona Crispie[1,2], Orla O'Sullivan[1,2], Laura Finnegan[1,2], Marcus J. Claesson[2,3] and Paul D. Cotter[1,2*]



Importance of databases

- Some are better curated than others
- Be consistent
- Update regularly
- Always put date of homology search in manuscripts

Ecological Diversity

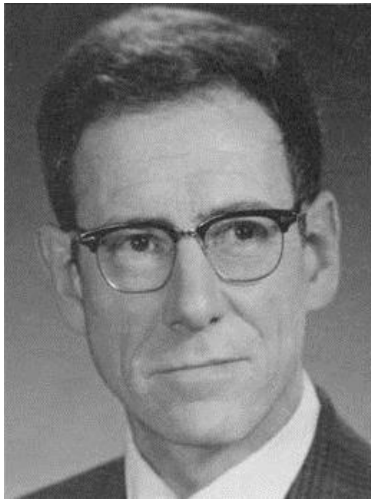
- Great place to begin getting to know your data

- **Alpha Diversity**

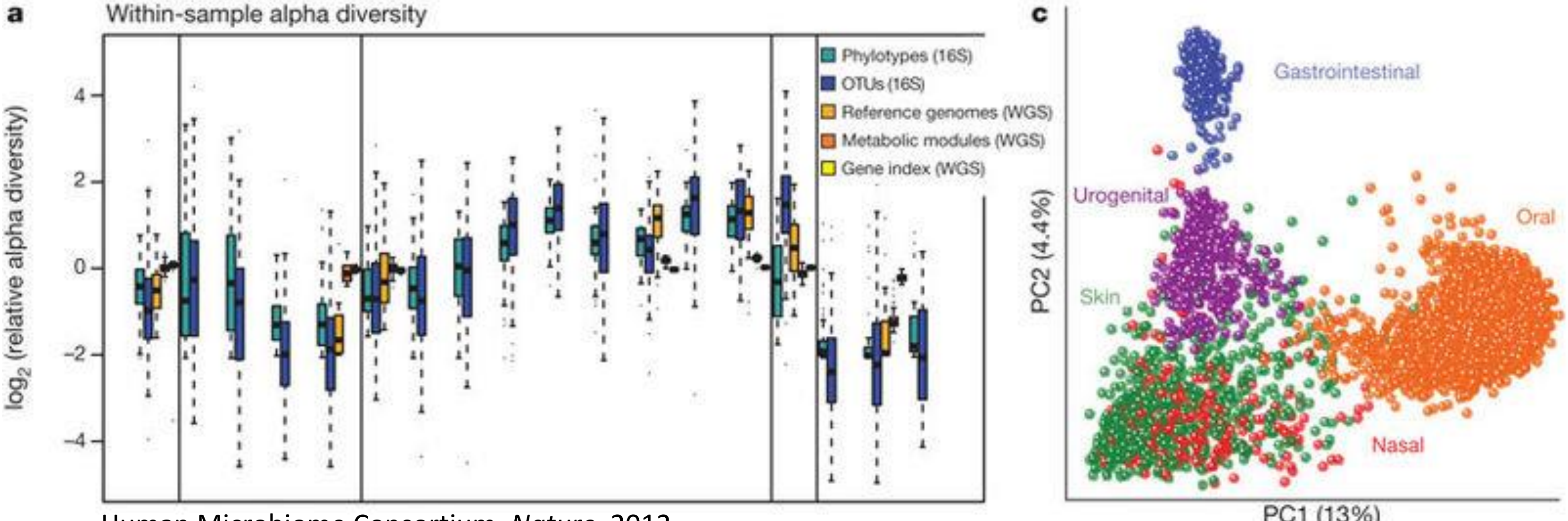
  -Biodiversity within a sample or community

- **Beta Diversity**

  -Difference between samples or communities



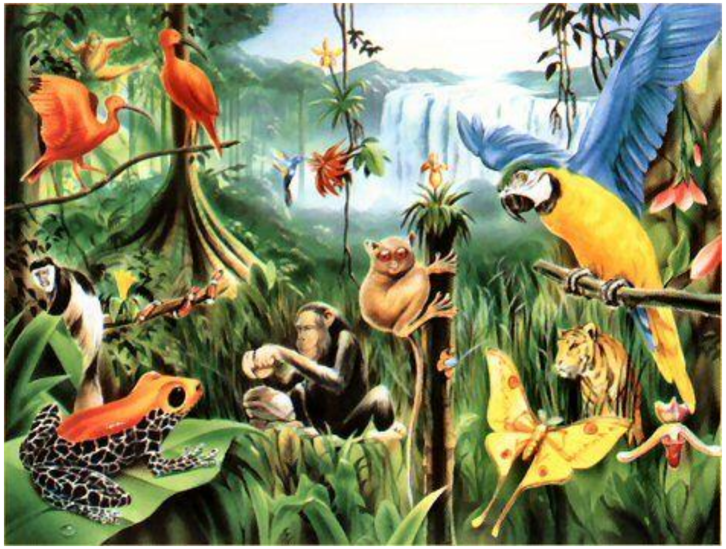
Robert Whittaker



Human Microbiome Consortium, *Nature*, 2012

Alpha Diversity

- Biodiversity within a sample or community



High Diversity

Low Diversity

Alpha Diversity Metrics

Observed species: number of different sequences in a sample

Phylogenetic diversity: incorporates phylogenetic difference between species. It is defined and calculated as "the sum of the lengths of all those branches that are members of the corresponding minimum spanning path".

Simpson : measure the degree of concentration when individuals are classified into types. So how much coverage is there

Chao1: (richness) Species richness is simply a count of species; it does not take into account the abundances of the species or their relative abundance distributions

Shannon: (Evenness) takes into account both abundance and evenness of species present in the community; species evenness quantifies how equal the abundances of the species are.

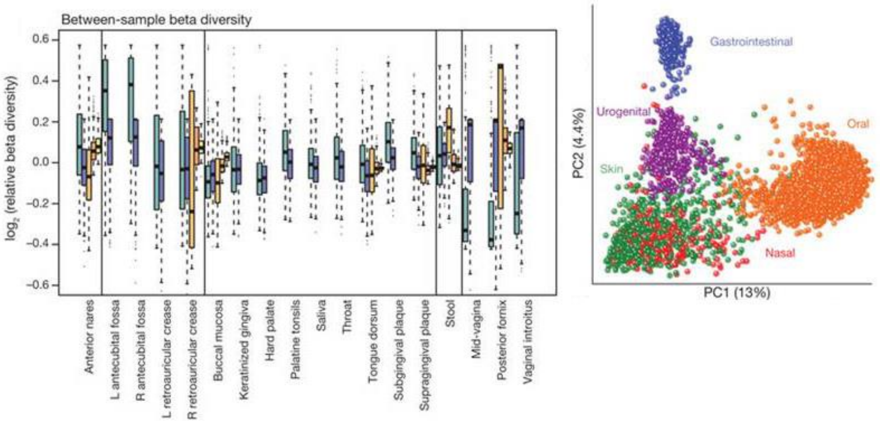
Beta Diversity

- Calculate the distance between a pair of samples
- Build up a distance matrix
- Matrix visualised in number of ways e.g.; network, PCoA, UPGMA tree
- Numerous metrics used to estimate distance

➢ e.g. – **Unifrac (dissimilarity measure)**: measure phylogenetic distance between sets of OTUs in a tree

<u>weighted Unifrac </u>takes into account relative abundance of OTUs
<u>unweighted Unifrac </u>no relative abundance.

- **bray curtis dissimilarity**: compares counts of OTUs between samples taking relative abundances into account



Human Microbiome Project Consortium, *Nature*, 2012

teagasc
Agriculture and Food Development Authority

Beta Diversity metrics in QIIME

- abund_jaccard
- binary_chisq
- binary_chord
- binary_euclidean
- binary_hamming
- binary_jaccard
- binary_lennon
- binary_ochiai
- binary_otu_gain
- binary_pearson
- binary_sorensen_dice
- bray_curtis
- bray_curtis_faith
- bray_curtis_magurran
- canberra
- chisq

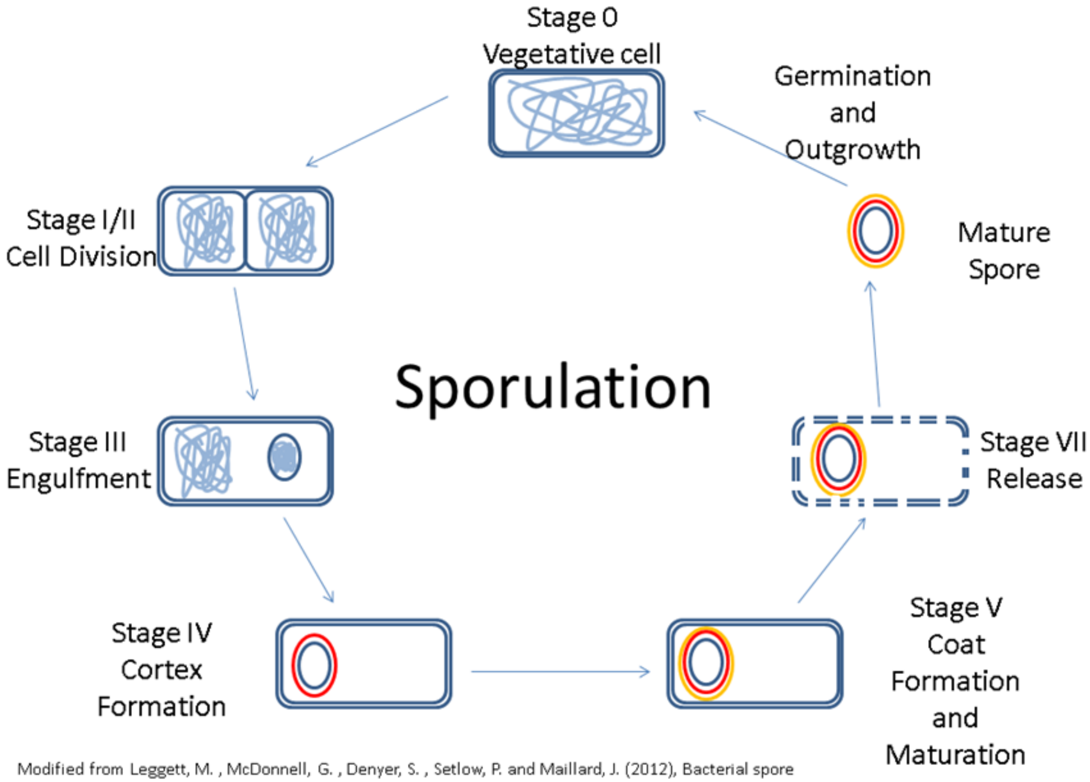
- chisq
- chord
- euclidean
- gower
- hellinger
- kulczynski
- manhattan
- morisita_horn
- pearson
- soergel
- spearman_approx
- specprof
- unifrac
- unifrac_g
- unifrac_g_full_tree
- unweighted_unifrac

- unweighted_unifrac_full_tree
- weighted_normalized_unifrac
- weighted_unifrac

- Binomial
- Mountfor
- Raup
- Cao
- Minkowski
- G-Unifrac
- Unifrac-VAW
- DPCoA
- JSD

Spore Detection in the Food Chain

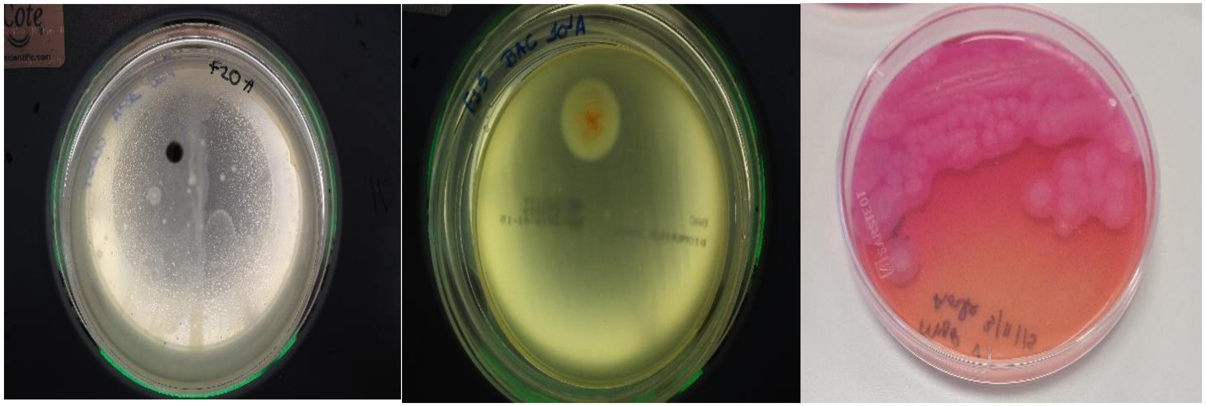
- Sporulate and remain dormant during processing, and equipment cleaning
- Germinate and proliferate in favourable conditions (i.e moisture and heat)
- Pathogenic?
- Associated with poor hygiene

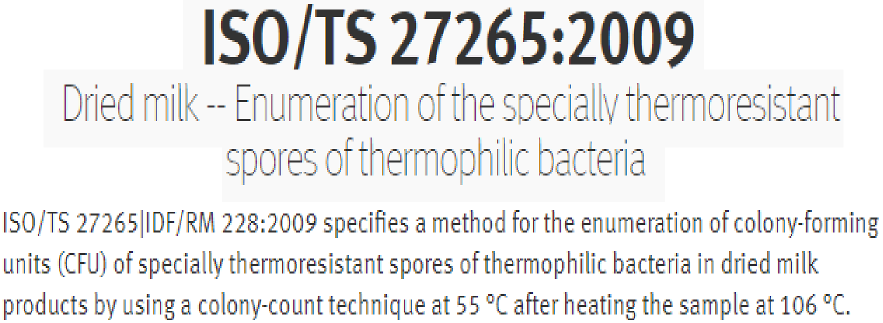


Modified from Leggett, M. , McDonnell, G. , Denyer, S. , Setlow, P. and Maillard, J. (2012), Bacterial spore structures and their protective role in biocide resistance. Journal of Applied Microbiology. 113: 485-498.

Traditional detection methods

- Phenotypic assays
- Plate counts



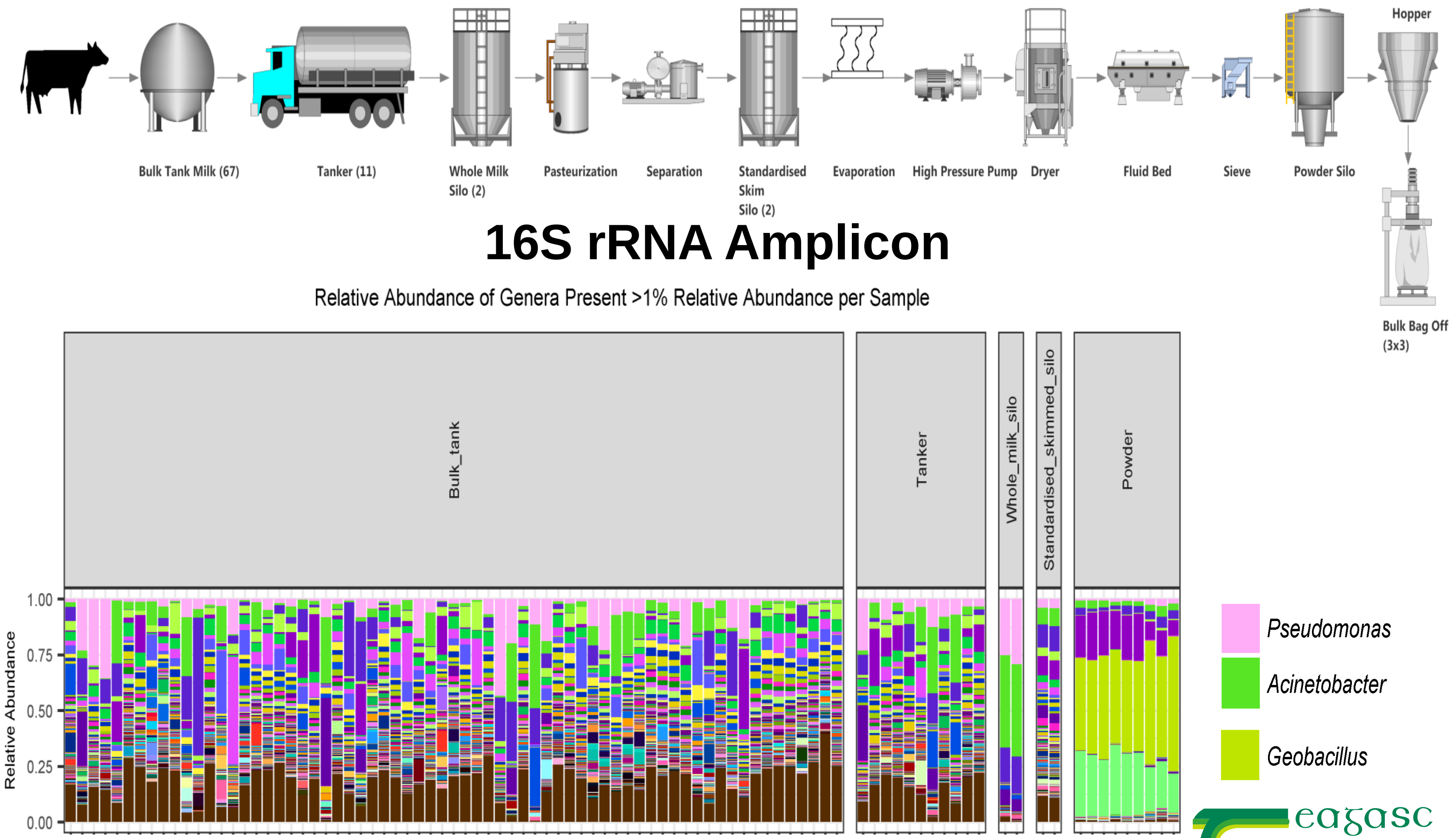


ISO/TS 27265:2009

Dried milk -- Enumeration of the specially thermoresistant spores of thermophilic bacteria

ISO/TS 27265|IDF/RM 228:2009 specifies a method for the enumeration of colony-forming units (CFU) of specially thermoresistant spores of thermophilic bacteria in dried milk products by using a colony-count technique at 55 °C after heating the sample at 106 °C.

↑ Processing ↓microbiota?

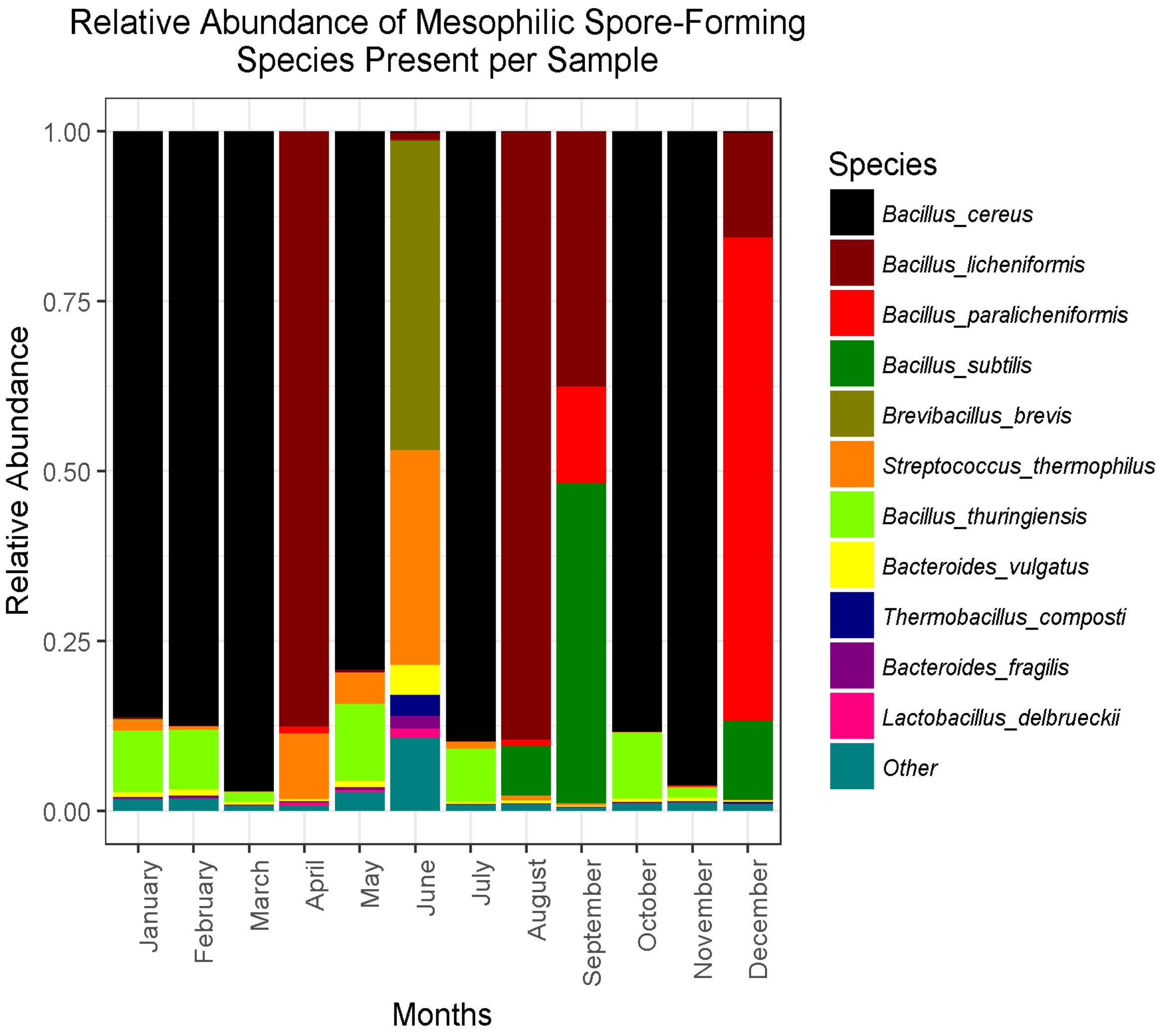


16S rRNA Amplicon

Relative Abundance of Genera Present >1% Relative Abundance per Sample

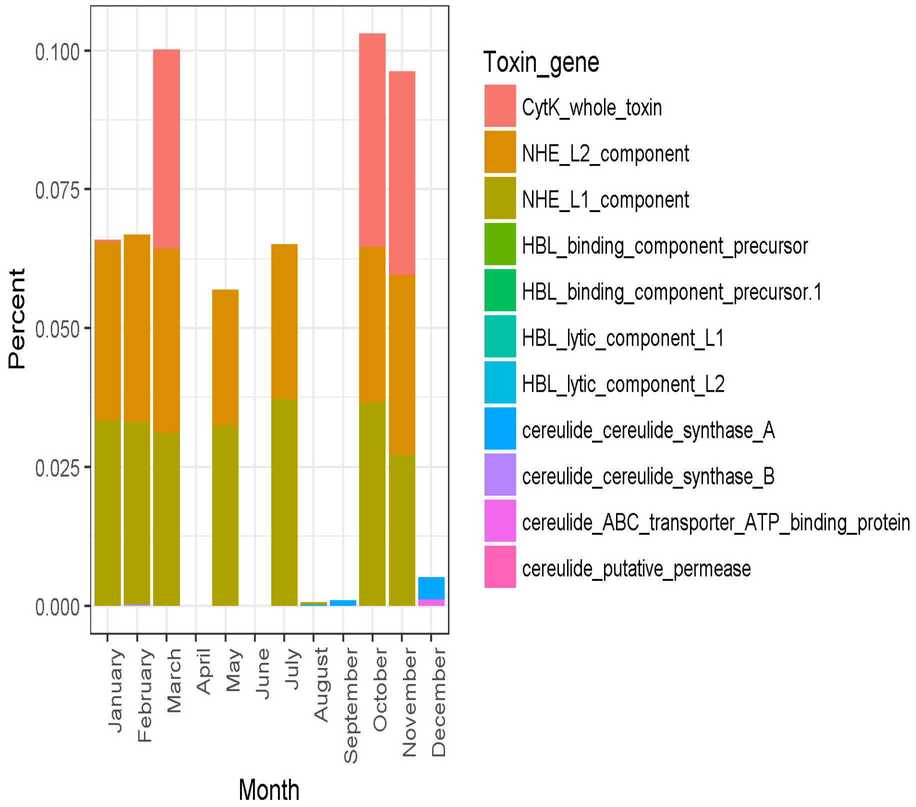
Teagasc
Agriculture and Food Development Authority

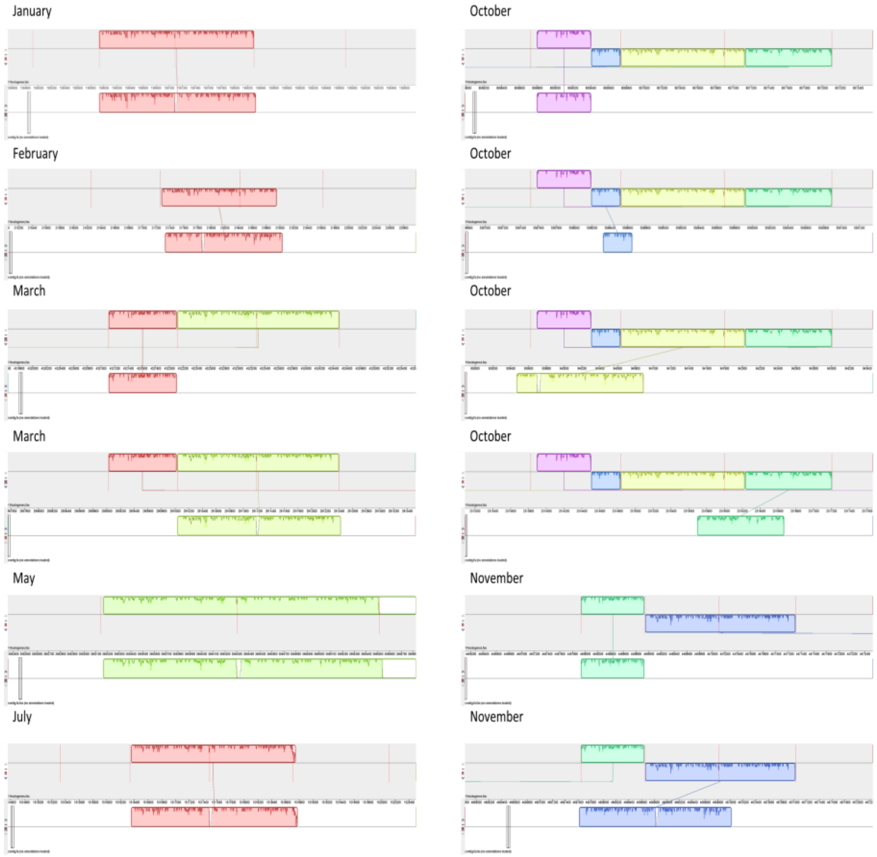
3 populations of mesophilic spore-formers identified



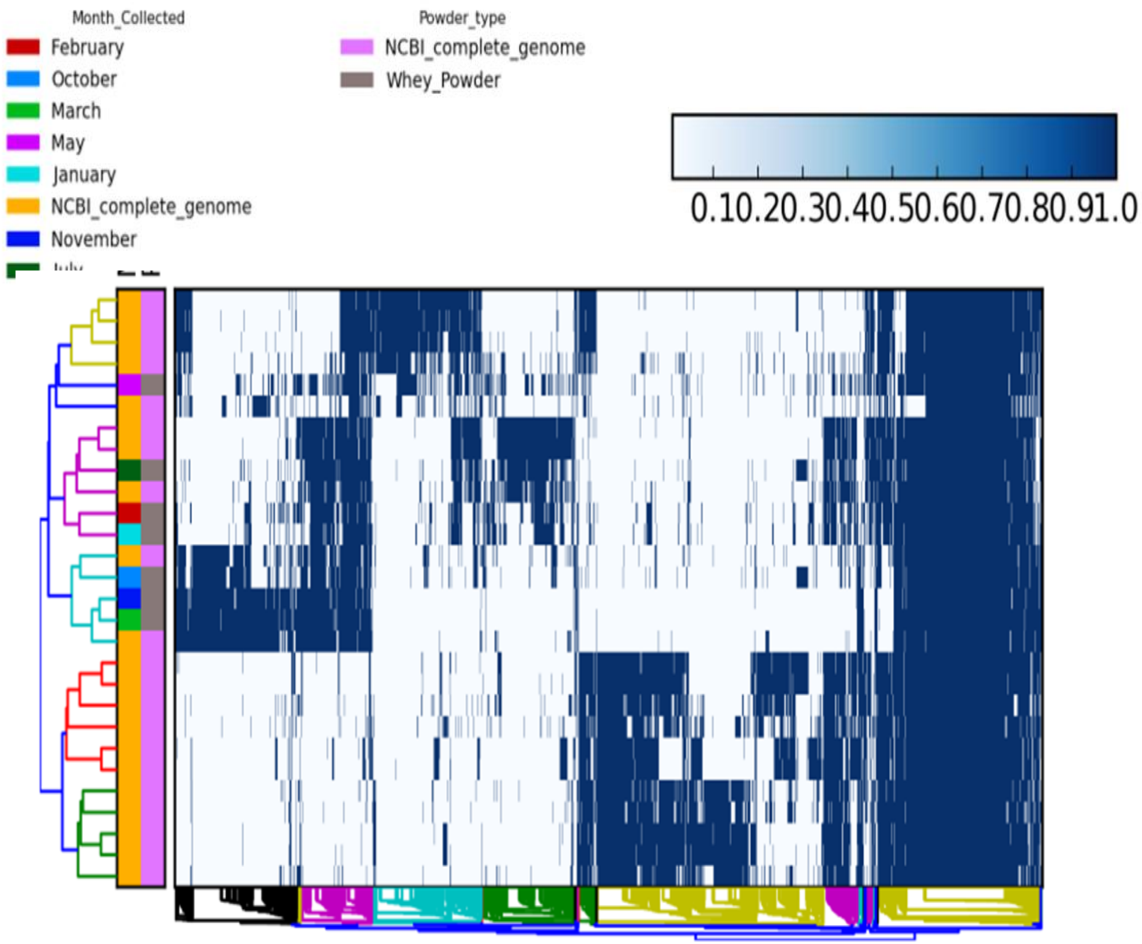
Putatively pathogenic *B. cereus* toxin gene analysis

Percent of Reads Attributed as *B. cereus* Toxin Genes





Putatively pathogenic *B. cereus* strain analysis



Conclusions

- Functional metagenomics has the potential to be used to delve deeper into the understanding of spore-formers in food-processing

Acknowledgements

Vision 1 group

apc Microbiome Ireland
Interfacing Food & Medicine

Science Foundation Ireland sfi For what's next

Teagasc
Agriculture and Food Development Authority