16S rRNA gene profiling

GastroPak Microbial Bioinformatics Workshop

DR CHRISTOPHER QUINCE

Earlham/Quadram Group Leader

University of East Anglia

Quadram Institute
Science Health Food Innovation

Earlham Institute
Decoding Living Systems

Overview

1) Shotgun metagenomics vs. 16S rRNA gene sequencing
2) Errors and biases in 16S rRNA sequencing
3) OTUs
4) Noise removal
5) Example of a 16S rRNA gene sequencing study

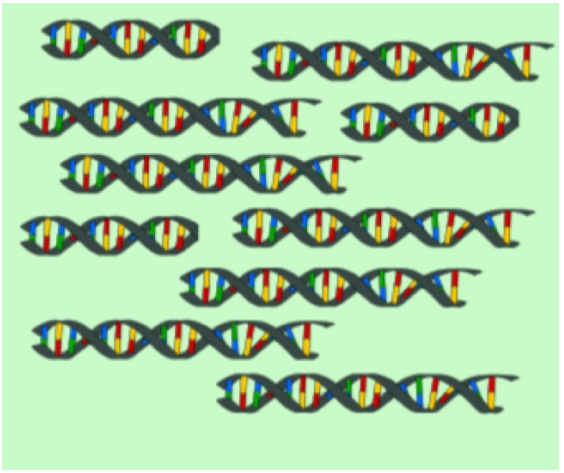
In vivo microbiome

Sample collection

In vivo sample

DNA extraction

Isolated DNA

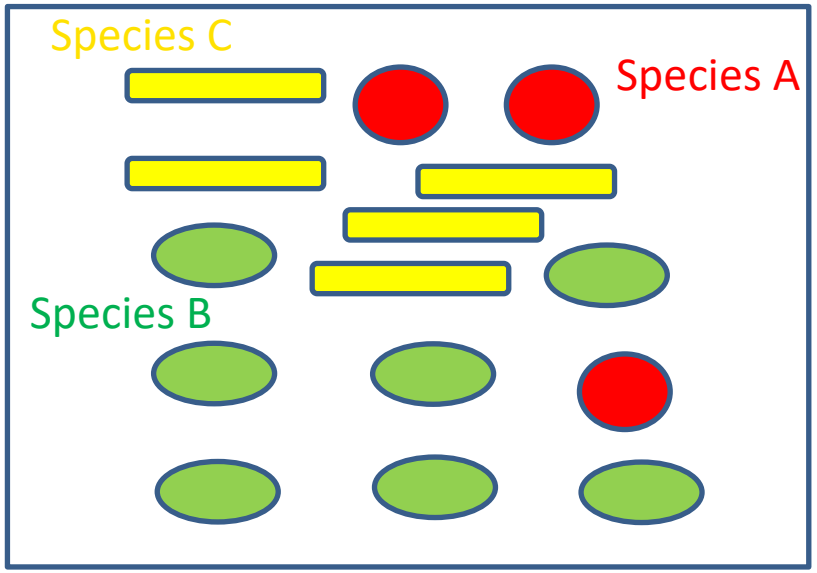


e.g. human gut,
soil, sediment

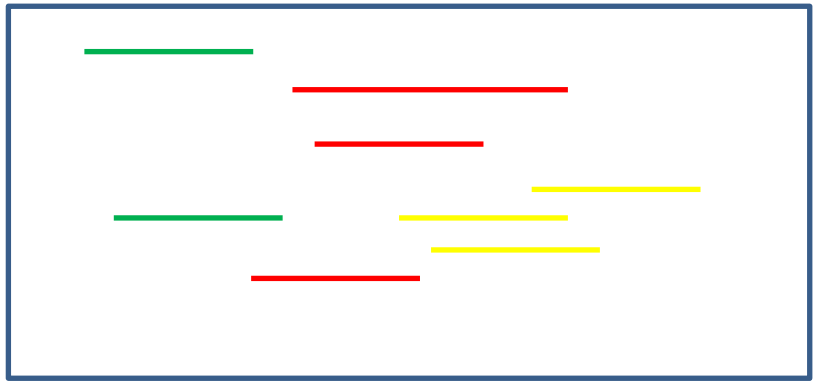
Microbiome sequencing vs. genomics

- DNA is obtained from the entire community
- Community is comprised of different species with different abundances
- DNA contains multiple genomes present in different proportions
- Variation within those species genomes

Community

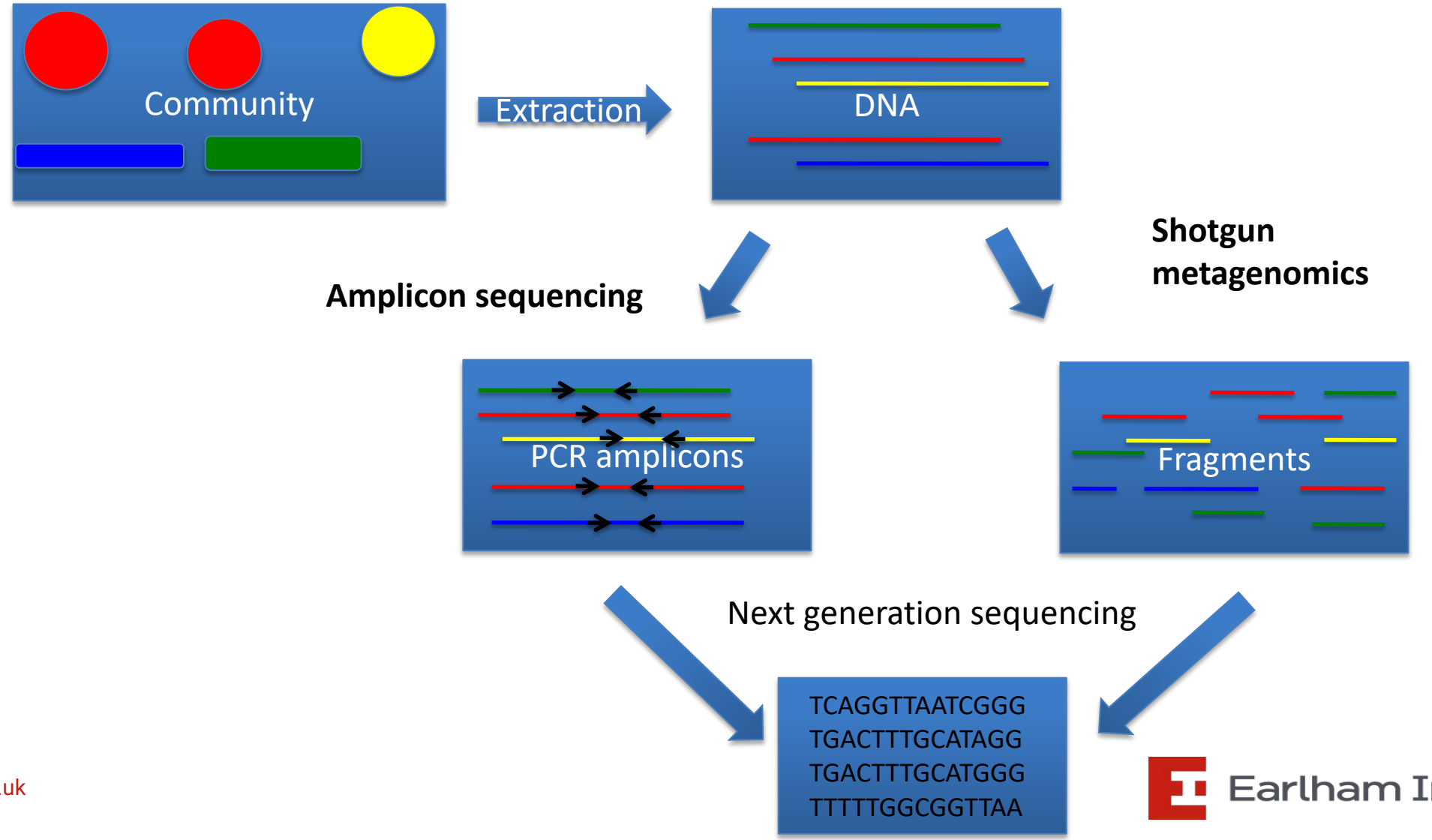


In DNA extract what will be the DNA proportions?



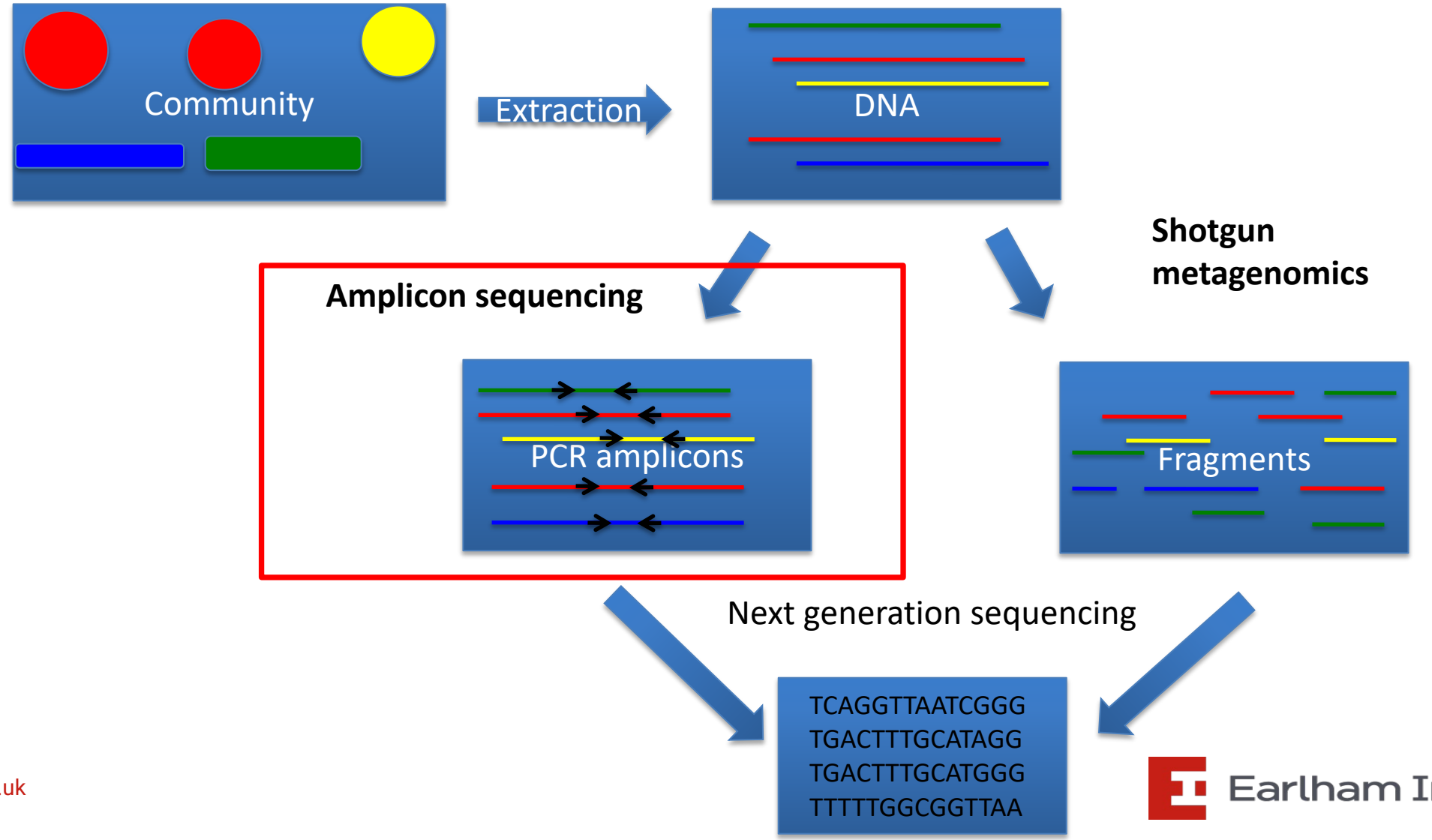
Species A: 5/15 = 33.3%
Species B: 7/15 = 46.7%
Species C: 3/15 = 20.0%

Amplicons vs. shotgun metagenomics



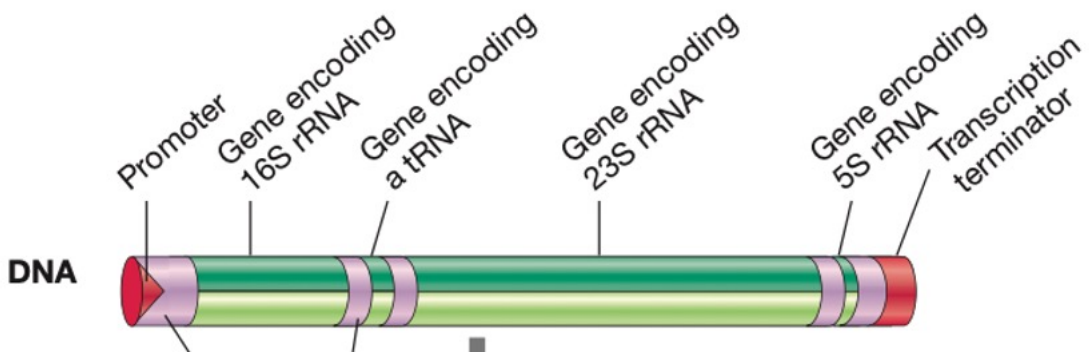
Earlham Institute

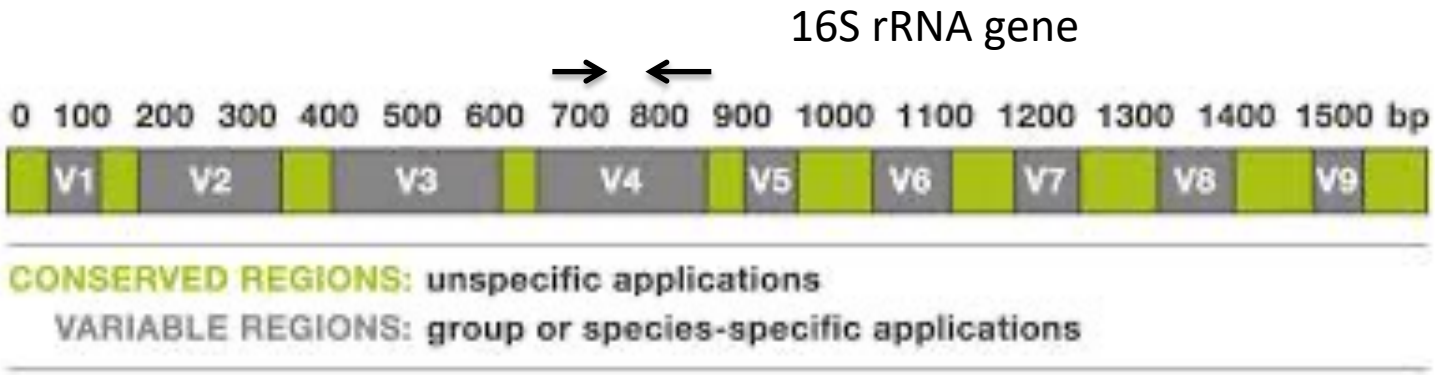
Amplicons vs. shotgun metagenomics

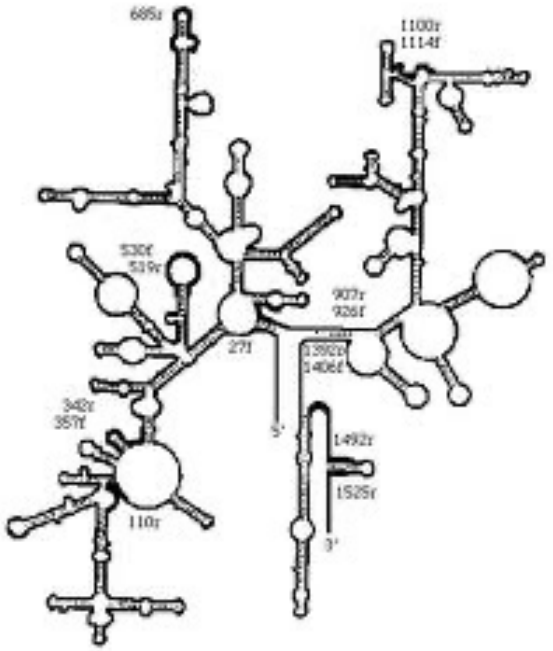


Earlham Institute

16S rRNA gene sequencing from a community





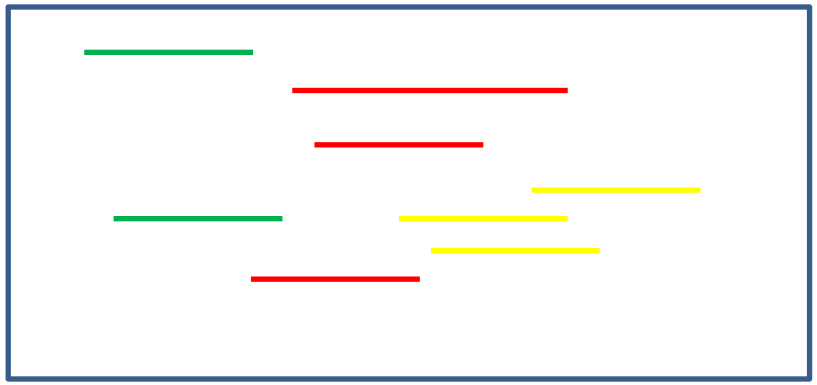


- Design PCR primers that are universal and target a subset of variable regions

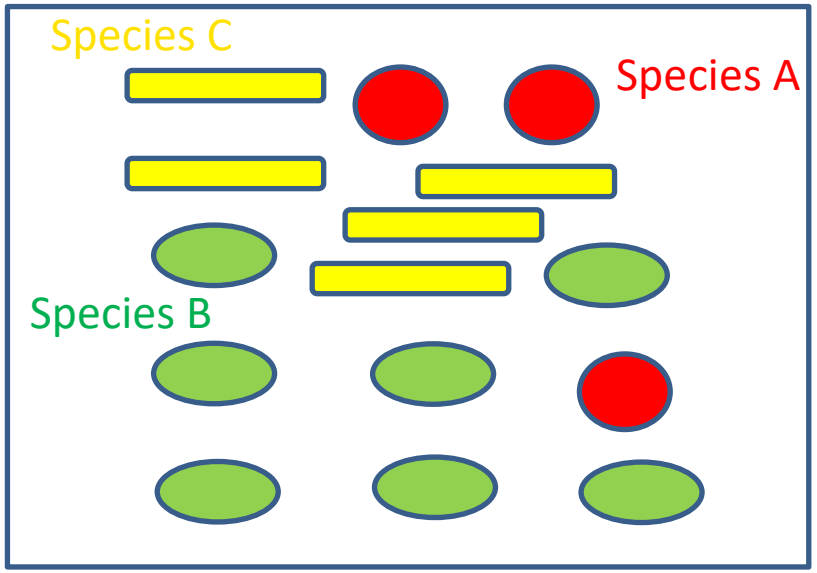
Earlham Institute

16S rRNA gene sequencing

In DNA extract what will be the 16S rRNA gene proportions?



Community



Species A: 5/15 = 33.3%
Species B: 7/15 = 46.7%
Species C: 3/15 = 20.0%

Have to consider variable rRNA operon number e.g. E coli has seven

In PCR product what will be the 16S rRNA gene proportions?

In sequenced data set?

› Elife. 2019 Sep 10;8:e46923. doi: 10.7554/eLife.46923.

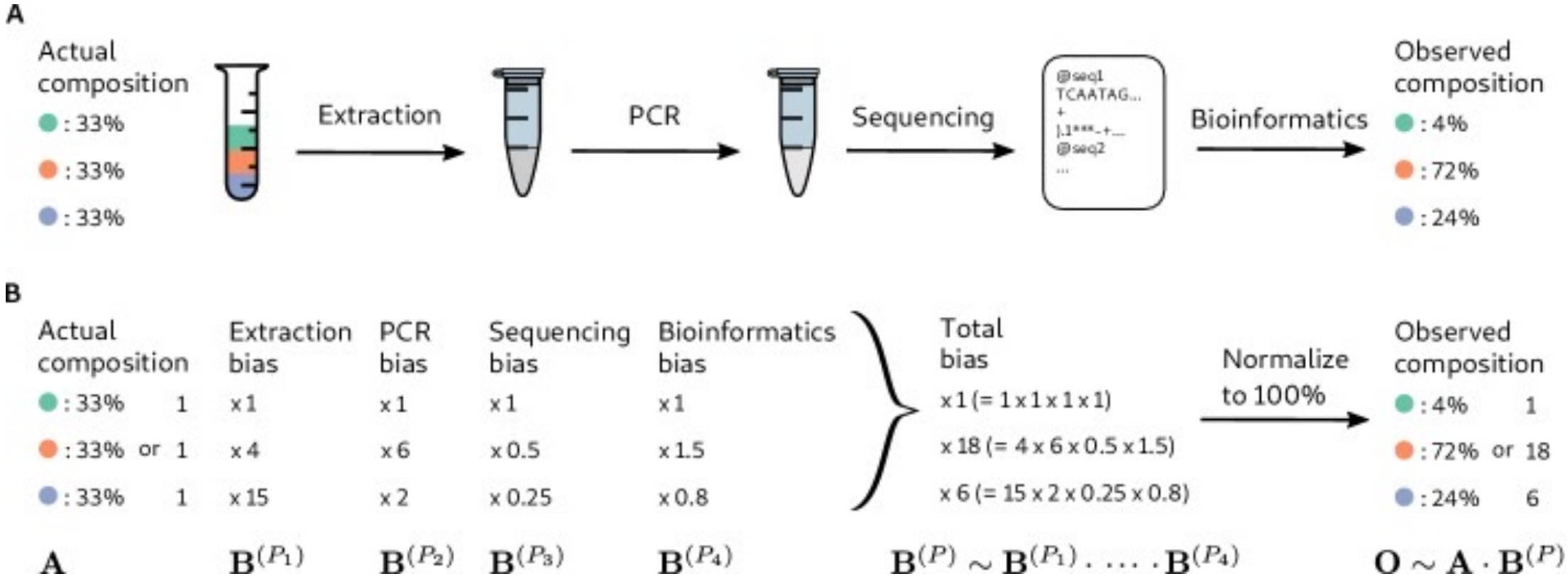
Consistent and correctable bias in metagenomic sequencing experiments

Michael R McLaren [1], Amy D Willis [2], Benjamin J Callahan [1 3]

Affiliations + expand

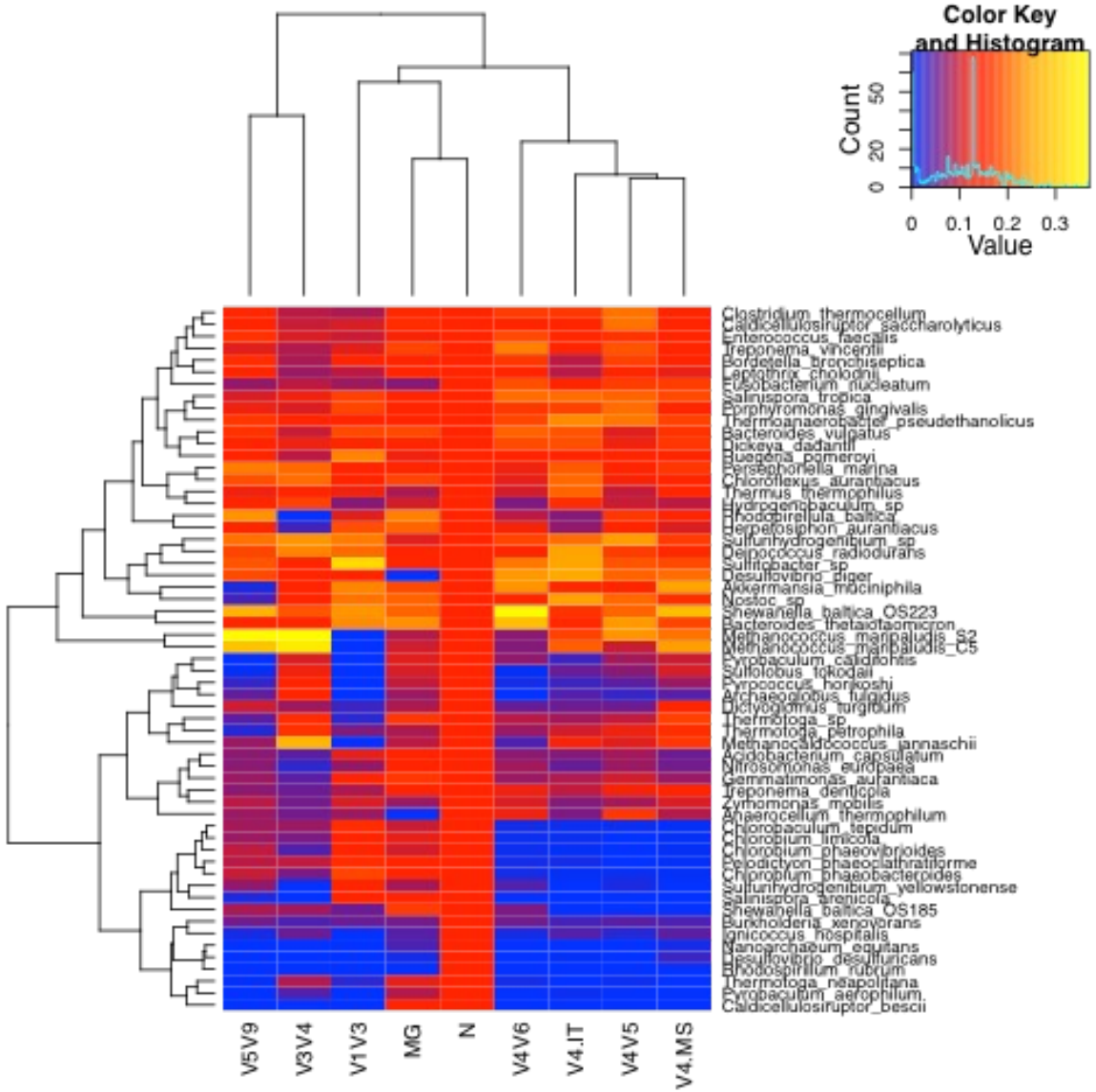
PMID: 31502536 PMCID: PMC6739870 DOI: 10.7554/eLife.46923

Free PMC article



Sources of error in 16S rRNA sequenced amplicons

- PCR bias represents the differential amplification of sequences – distorts relative abundances – no such thing as universal primer
- Mock community evaluations



Research Article | Open Access | Published: 14 January 2016

A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling

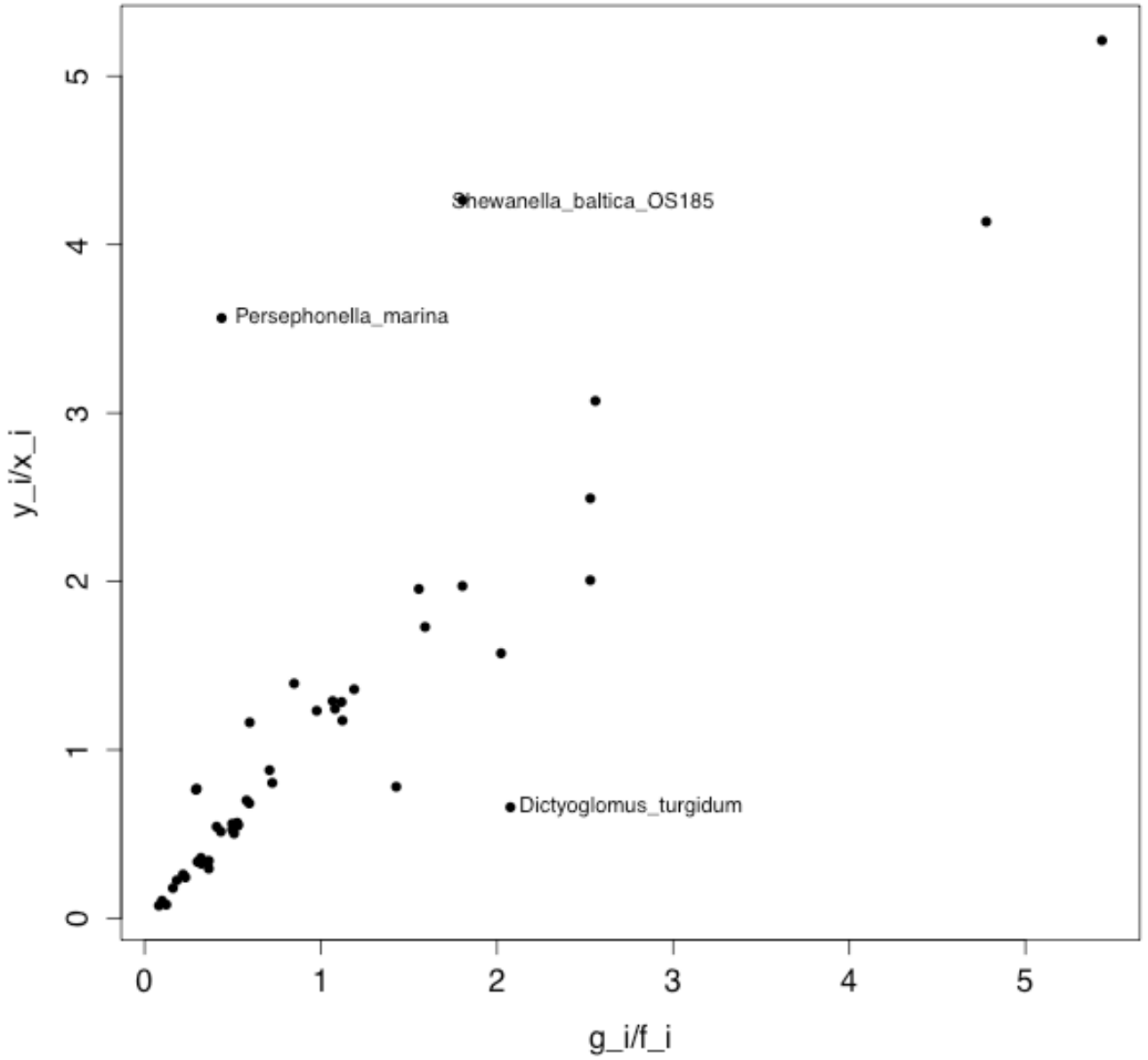
Rosalinda D'Amore, Umer Zeeshan Ijaz, Melanie Schirmer, John G. Kenny, Richard Gregory, Alistair C. Darby, Migun Shakya, Mircea Podar, Christopher Quince & Neil Hall

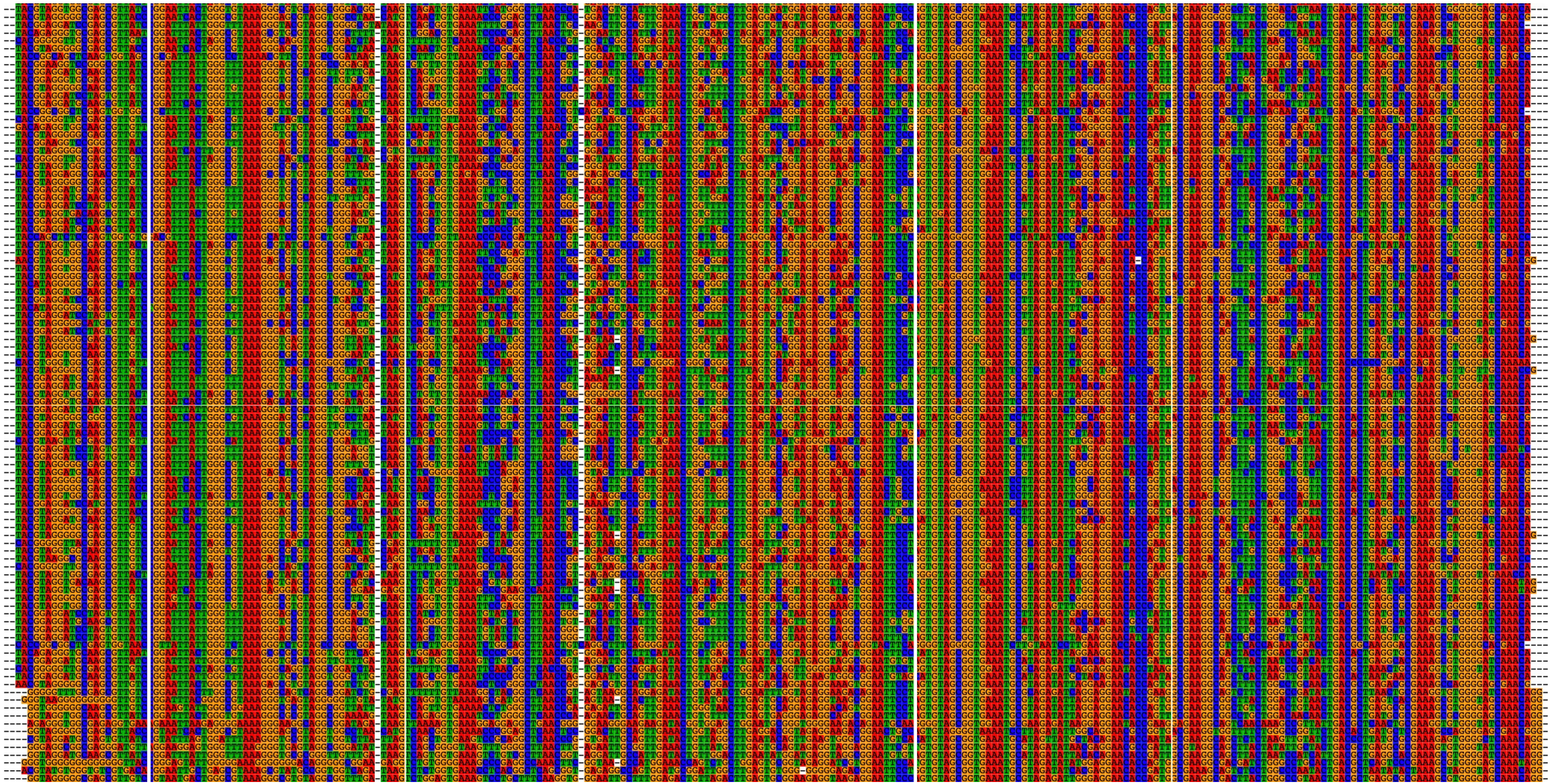
BMC Genomics **17**, Article number: 55 (2016) | Cite this article

16k Accesses | **229** Citations | **45** Altmetric | Metrics

Biased but still quantitative…

Coefficient: 1.00560
Adjusted R-squared: 0.8333
p-value: < 2.2e-16





Taxonomic classification of 16S rRNA sequences

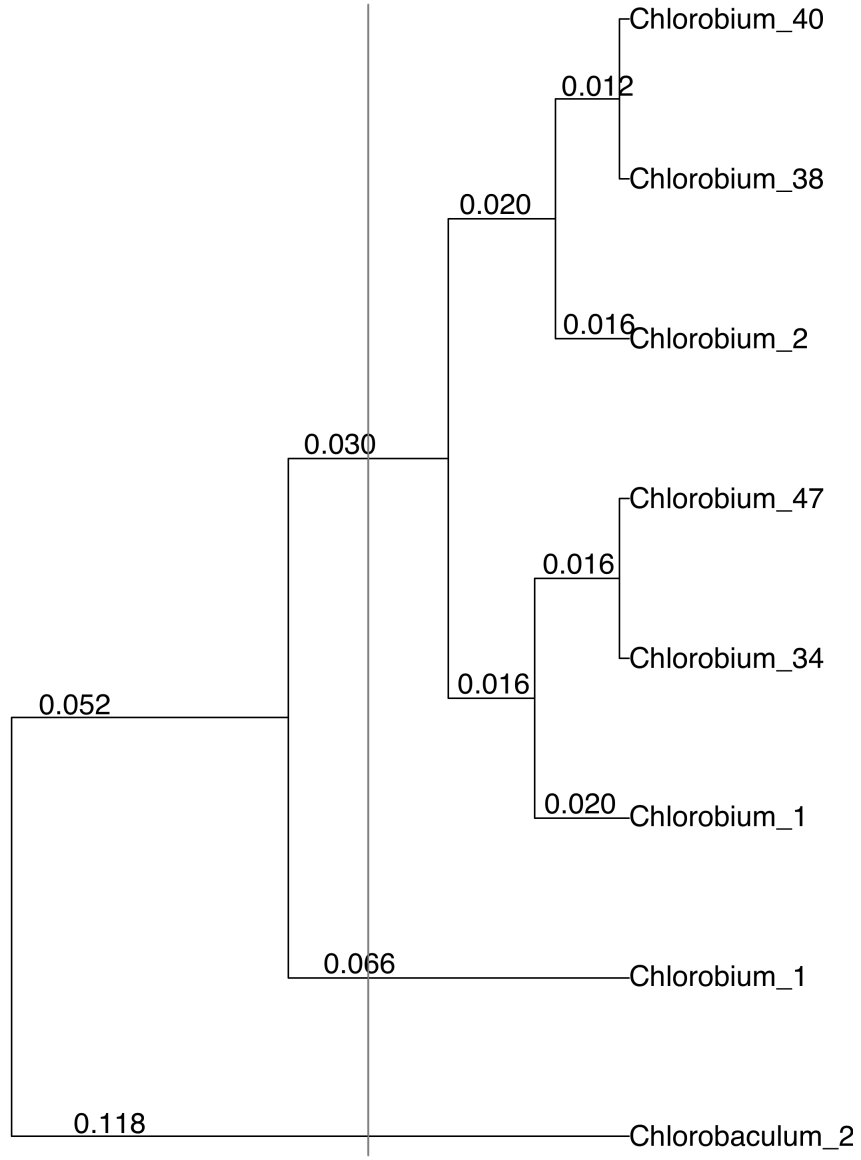
- Homology search against database e.g. vsearch
- Global alignment i.e. Needleman-Wunsch algorithm
- Once sequences are aligned metrics are calculated to indicate similarity to reference sequences
  - Edit distance

    ACTGCTTTAGGGGG -> database
    ACT- CTTAAGGGGT -> query

    Edit distance = 3
  - E value: describes the number of hits one can "expect" to see by chance when searching a database of a particular size
- Typically top N hits are returned
- Find sequence with closest match (%id to query) and assign
- Lowest common ancestor amongst top hits
- Problem most taxa not in database - true circa 2014 ☺

Operational taxonomic units

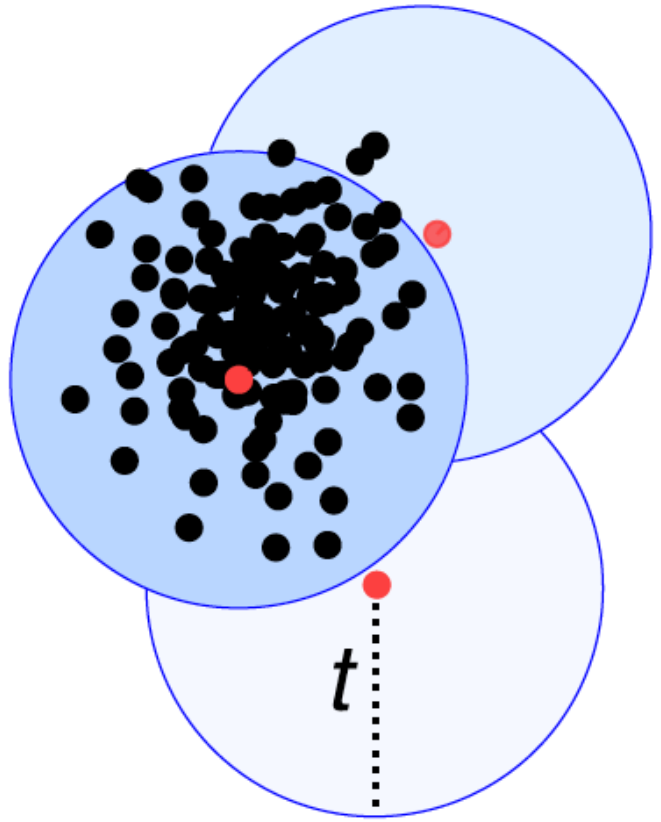
- OTUs *de novo* sequence clusters
- Attempt to assign sequences to meaningful biological entities without a reference database
- Simplest strategy hierarchical clustering
  1) Align sequences
  2) Calculate distance matrix
  3) Hierarchical clustering (complete linkage or average linkage – UPGMA)
  4) OTUs defined at a specified cut-off (e.g. 3% species)
- Alignment errors can be avoided by calculating exact pairwise sequence distances



Earlham Institute

One pass centroid clustering

- N^2 hierarchical algorithm too slow for MiSeq data
- Instead one pass centroid clustering as implemented in U-parse or V-search
  - Compare each sequence in turn to centroids if distance less than threshold *t* assign to that cluster
  - If not sequence forms new centroid
- Uses 3% dissimilarity cut-off and discard singletons



The ‘rare biosphere’

- I began working on microbiome data in 2007
- Initial 16S rRNA amplicon studies based on GS20 454 pyrosequencing were reporting huge levels of diversity
- Supported earlier studies based on DNA reassociation

Microbial Population Structures in the Deep Marine Biosphere

Julie A. Huber,[1]* David B. Mark Welch,[1] Hilary G. Morrison,[1] Susan M. Huse,[1] Phillip R. Neal,[1] David A. Butterfield,[2] Mitchell L. Sogin[1]

The analytical power of environmental DNA sequences for modeling microbial ecosystems depends on accurate assessments of population structure, including diversity (richness) and relative abundance (evenness). We investigated both aspects of population structure for microbial communities at two neighboring hydrothermal vents by examining the sequences of more than 900,000 microbial small-subunit ribosomal RNA amplicons. The two vent communities have different population structures that reflect local geochemical regimes. Descriptions of archaeal diversity were nearly exhaustive, but despite collecting an unparalleled number of sequences, statistical analyses indicated additional bacterial diversity at every taxonomic level. We predict that hundreds of thousands of sequences will be necessary to capture the vast diversity of microbial communities, and that different patterns of evenness for both high- and low-abundance taxa may be important in defining microbial ecosystem dynamics.

The interrogation of DNA from environmental samples has revealed new dimensions in microbial diversity and community-wide metabolic potential. The first analysis of a dozen polymerase chain reaction (PCR) amplicons of ribosomal RNA (rRNA) sequence from a mixed bacterioplankton population revealed the ubiquitous SAR11 cluster (*1*), and a recent environmental shotgun sequence survey of microbial communities in the surface ocean has identified 6.1 million predicted proteins (*2*, *3*). To realize the full potential of metagenomics for modeling energy and carbon flow, microbial biogeography, and the relationship between microbial diversity and ecosystem function, it is necessary to estimate both the richness and evenness of microbial population structures.

We used a tag sequencing strategy that combines the use of amplicons of the V6 hypervariable region of small-subunit (SSU) rRNA as proxies for the presence of individual phylotypes [operational taxonomic units (OTUs)] with massively parallel sequencing. Our goal was to provide assessments of microbial diversity, evenness, and community structure at a resolution two to three orders of magnitude greater than that afforded by cloning and capillary sequencing of longer SSU rRNA amplicons (*4*). We used this strategy to attempt an exhaustive characterization of the bacterial and archaeal diversity at two

REPORTS

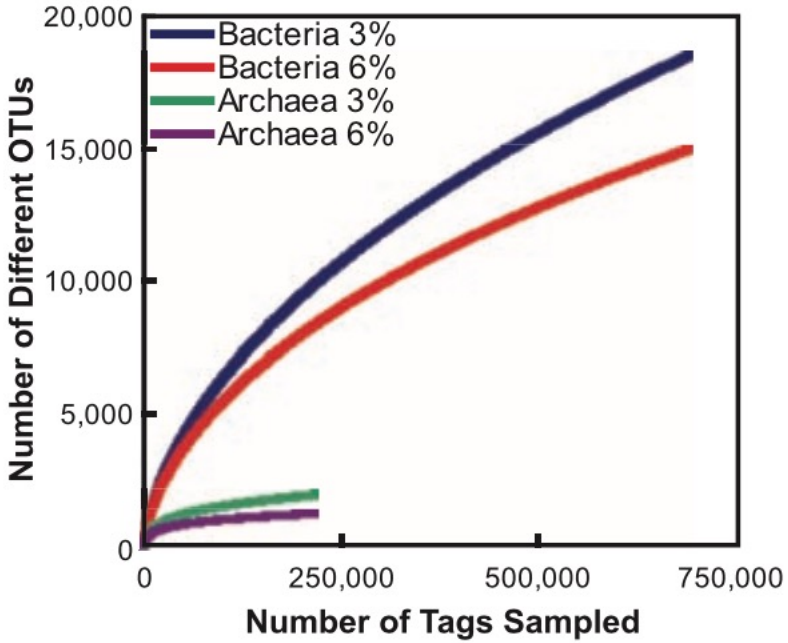
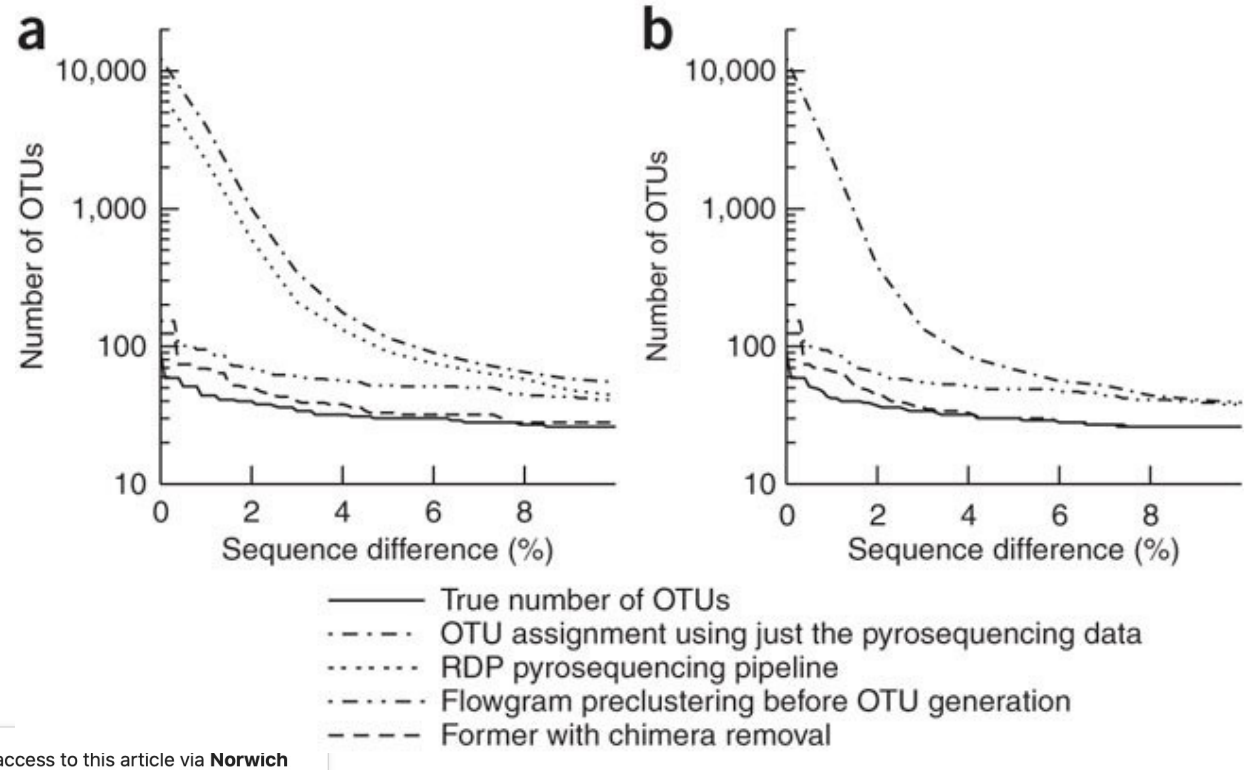


Fig. 2. Rarefaction curves for total bacterial and archaeal communities at the two sampling sites FS312 and FS396 at 3% and 6% difference levels.

Earlham Institute

‘Rare biasphere?’

- Developed PyroNoise algorithm to allow read denoising and accurate diversity estimation
- First paper to quantify diversity inflation
- Ideas in PyroNoise helped lead to current pipelines such as DADA2
- Follow up paper AmpliconNoise included the first de novo chimera removal algorithms



Published: 09 August 2009

Accurate determination of microbial diversity from 454 pyrosequencing data

Christopher Quince, Anders Lanzén, Thomas P Curtis, Russell J Davenport, Neil Hall, Ian M Head, L Fiona Read & William T Sloan

Nature Methods **6**, 639–641 (2009) | Cite this article

Abstract

We present an algorithm, PyroNoise, that clusters the flowgrams of 454 pyrosequencing reads using a distance measure that models sequencing noise. This infers the true sequences in a collection of amplicons. We pyrosequenced a known mixture of microbial 16S rDNA sequences extracted from a lake and found that without noise reduction the number of operational taxonomic units is overestimated but using PyroNoise it can be accurately calculated.

Associated Content

The 'rare biosphere': a reality check

Jens Reeder & Rob Knight

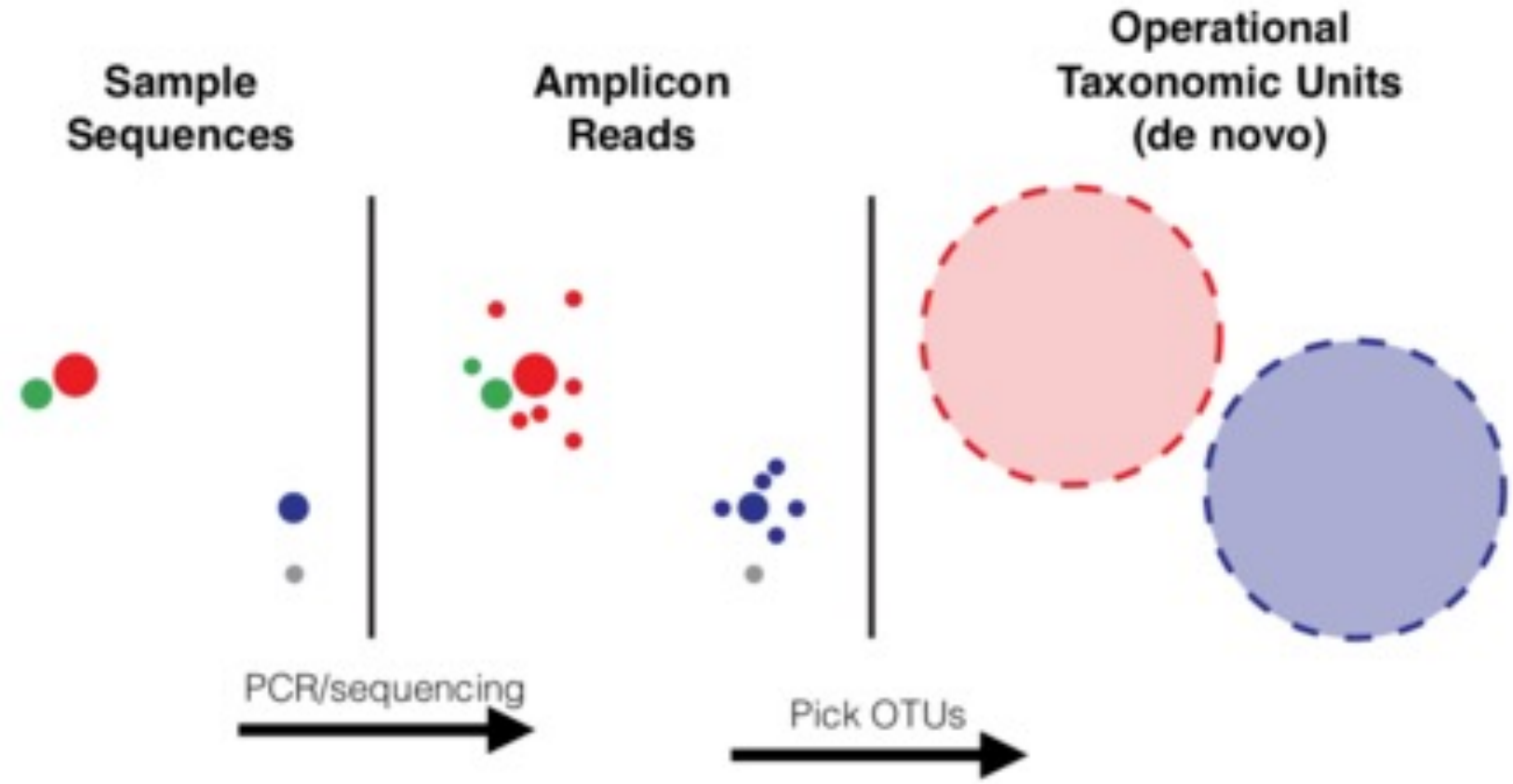
Nature Methods | **News & Views** | 01 Sept 2009

Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates

Victor Kunin, Anna Engelbrektson, Howard Ochman, Philip Hugenholtz

First published: 29 December 2009 | https://doi.org/10.1111/j.1462-2920.2009.02051.x | Citations: 784

Errors both sequencing and PCR noise generate spurious OTUs



DADA2 error correction

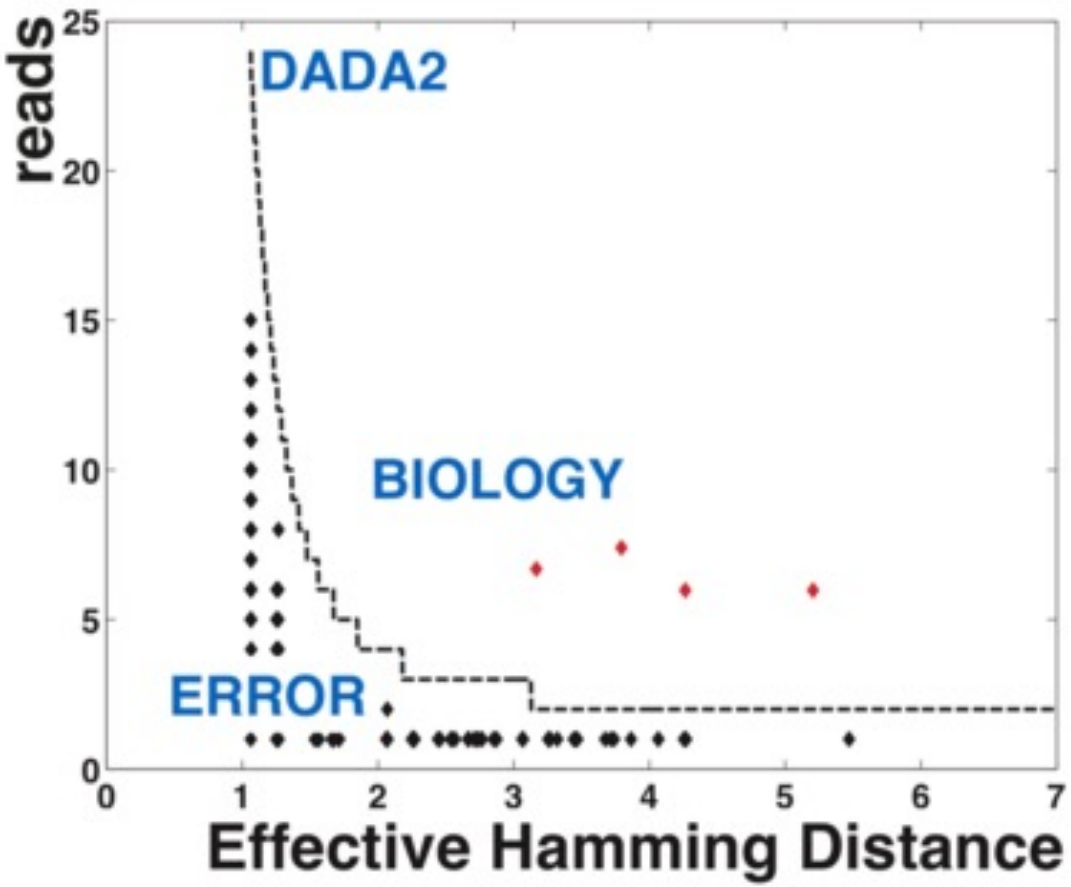
- Divisive Amplicon Denoising Algorithm 2 (DADA2)
- Contains Poisson error model for observed read to be generated from a sequence with a given abundance as noise
- If this is unlikely observed read is judged to be a true sequence
- Ignores singletons

$$p_A(j \to i) = \frac{1}{1 - \rho_{pois}(n_j \lambda_{ji}, 0)} \sum_{a=a_i}^{\infty} \rho_{pois}(n_j \lambda_{ji}, a)$$

Nat Methods. Author manuscript; available in PMC 2016 Nov 23.
Published in final edited form as:
Nat Methods. 2016 Jul; 13(7): 581–583.
Published online 2016 May 23. doi: 10.1038/nmeth.3869
PMCID: PMC4927377
NIHMSID: NIHMS782534
PMID: 27214047

DADA2: High resolution sample inference from Illumina amplicon data

Benjamin J Callahan,[1,*] Paul J McMurdie,[2] Michael J Rosen,[3] Andrew W Han,[2] Amy Jo A Johnson,[2] and Susan P Holmes[1]



Amplicon sequence variants

ISME J. 2017 Dec; 11(12): 2639–2643.
Published online 2017 Jul 21. doi: 10.1038/ismej.2017.119
PMCID: PMC5702726
PMID: 28731476

ISME J

Exact sequence variants should replace operational taxonomic units in marker-gene data analysis

Benjamin J Callahan,[1,*] Paul J McMurdie,[2] and Susan P Holmes[3]

Author information · Article notes · Copyright and License information Disclaimer

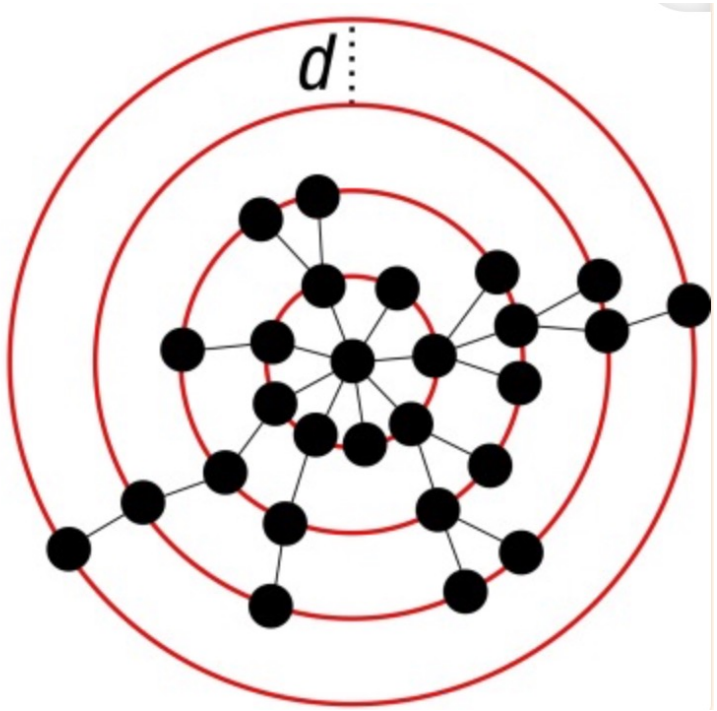
Abstract

Go to:

Earlham Institute

16S rRNA gene bioinformatics - additional topics

- Mock community evaluation of noise, error biases
- Variety of OTU clustering algorithms, swarm, ecological clustering
- Functional inference e.g. PICrust
- Integrated pipelines QIIME or LOTUS2

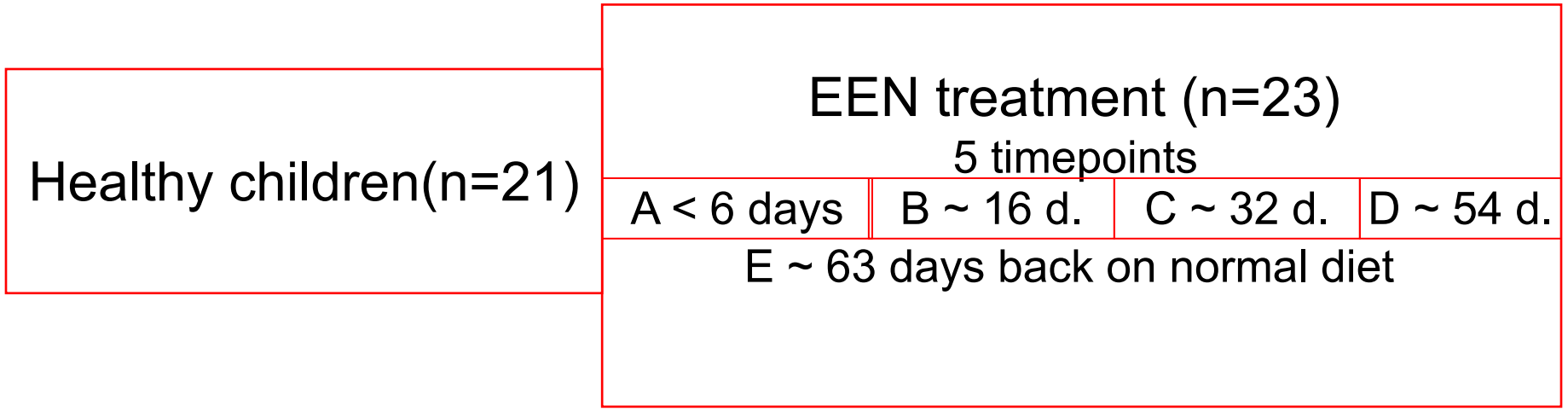


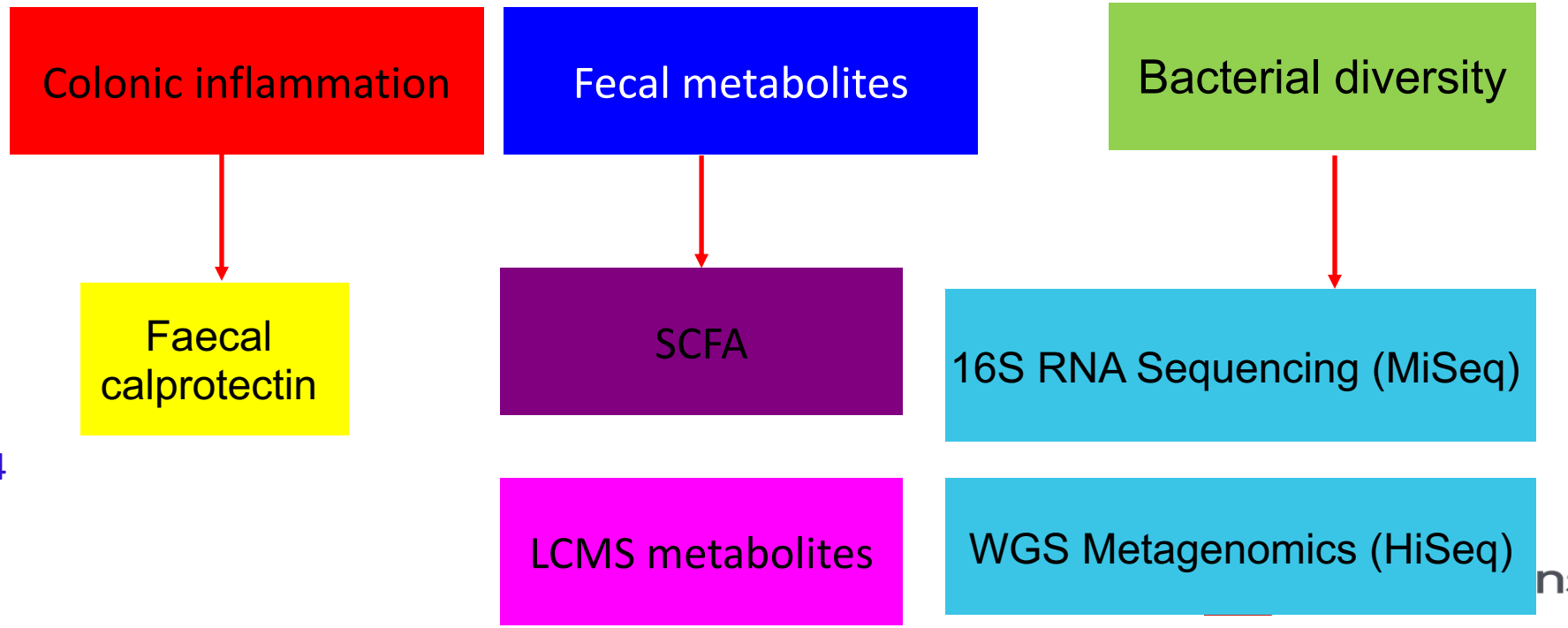
Crohn’s disease

- Chronic inflammatory condition of the GI tract disease that is increasing in developed world
- A gut microbiota, characteristic of Crohn's disease (CD), has been described from cross-sectional studies e.g. Gevers et al. Cell Host Microbe (2014):
  - CD associated: Veilonella, Escherichia, Haeomophilus
  - H associated: Ruminococcus, Faecalibacterium, Coprococcus
- Instead focussed on longitudinal study of pediatric Crohn’s with Exclusive Enteral Nutrition (EEN) – run by Prof. Konstantinos Gerasimidis:
  - Duration 8 weeks
  - **No other food**
  - Achieves clinical remission in large majority (up to 80%)

University of Glasgow

CICRA Study design

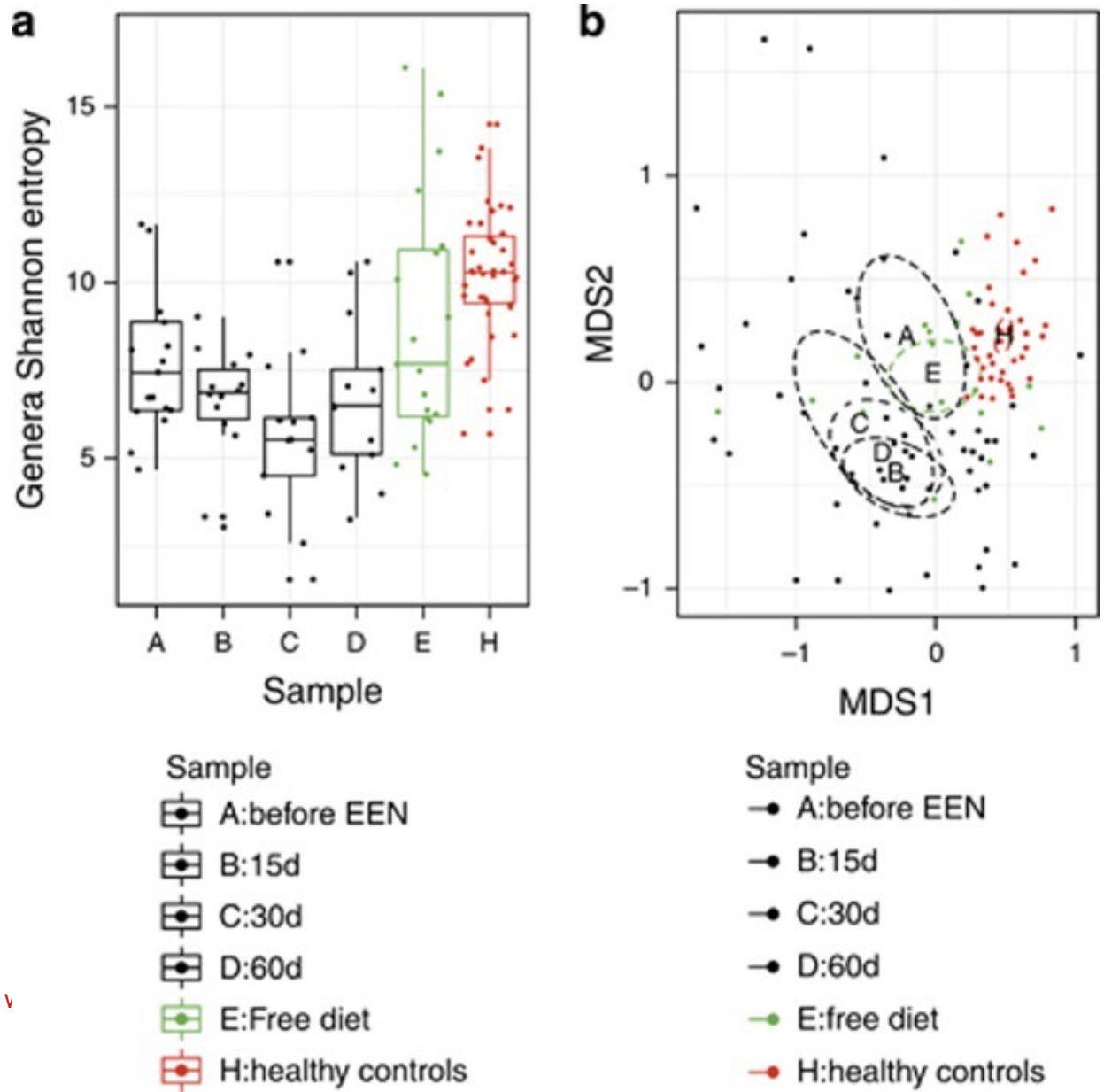




Decreased Butyrate and *Faecalibacterium prausnitzii* (qPCR) with treatment (Gerasimidis et al. Inflamm Bowel Dis. 2014 May;20(5):861-71.)

nstitute

16S rRNA gene results



Am J Gastroenterol. 2015 Dec; 110(12): 1718–1729.
Published online 2015 Nov 3. doi: 10.1038/ajg.2015.357
PMCID: PMC4697132
PMID: 26526081

Extensive Modulation of the Fecal Metagenome in Children With Crohn's Disease During Exclusive Enteral Nutrition

Christopher Quince, PhD, BSc,[1,8] Umer Zeeshan Ijaz, PhD, BSc,[2,8] Nick Loman, PhD, MBChB,[3] A Murat Eren, PhD, BSc,[4] Delphine Saulnier, PhD, BSc,[5] Julie Russell, BSc,[2] Sarah J Haig, PhD, BSc,[2] Szymon T Calus, BSc,[3] Joshua Quick, PhD, BSc,[3] Andrew Barclay, MD, MBChB,[6] Martin Bertz, MSc, BSc,[5] Michael Blaut, PhD, BSc,[5] Richard Hansen, PhD, MBChB,[6] Paraic McGrogan, MBChB,[6] Richard K Russell, PhD, MBChB,[6] Christine A Edwards, PhD, BSc,[7] and Konstantinos Gerasimidis, PhD, MSc, BSc[7,*]

Earlham Institute

Summary

- 16S rRNA sequencing is still an effective way to obtain the taxonomic profile of a community
- Despite noise and biases it is good for detecting changes in community structure and species biomarkers
- Other marker genes are possible e.g. rpoB

A Comparison of *rpoB* and 16S rRNA as Markers in Pyrosequencing Studies of Bacterial Diversity

Michiel Vos, Christopher Quince, Agata S. Pijl, Mattias de Hollander, George A. Kowalchuk

Published: February 15, 2012 • https://doi.org/10.1371/journal.pone.0030600