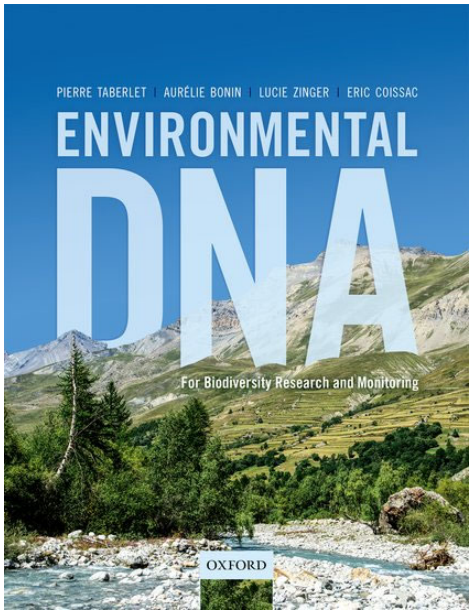
BIO9905MERG1 – Bioinformatics for Environmental Sequencing (DNA metabarcoding)

Welcome!

Introduction to DNA metabarcoding

- Explain terms
- Introduce key steps
- Introduce some literature
- More in-depth information in later talks

Pierre Taberlet



Some important / confusing terms

ASVs

amplicon sequencing

OTUs

DNA metabarcoding

metatranscriptomics

eDNA

rDNA phylotyping

metagenetics

metagenomics

microbiome analysis

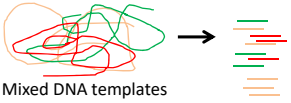
environmental sequencing

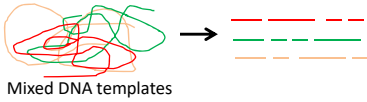
MOTUs

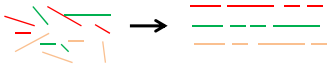
community profiling

marker gene analysis

Some important terms

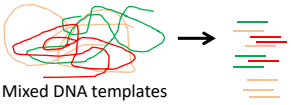
- DNA barcoding  Sequence variation in a single locus (e.g. ITS) in a single specimen
- Metabarcoding 

Sequence variation in a single locus (e.g. ITS) in a community
- Metagenomics 

Genome wide sequence variation in a community
- Metatranscriptomics 

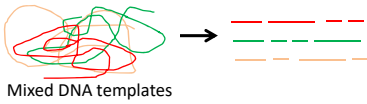
cDNA sequence variation in a community

Some important terms

- Metabarcoding 

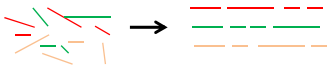
Sequence variation in a single locus (e.g. 16S)

Who are there?

- Metagenomics 

Genome wide sequence variation

Which genes (and who) are there?

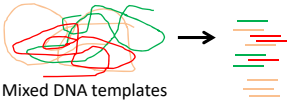
- Metatranscriptomics 

cDNA sequence variation

Who are active and doing what?

Some important terms

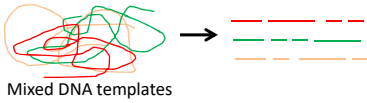
- Metabarcoding



Sequence variation in a single locus (e.g. 16S)

Who are there?

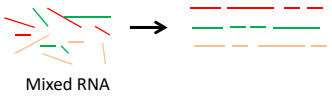
- Metagenomics



Genome wide sequence variation

Which genes (and who) are there?

- Metatranscriptomics



cDNA sequence variation

Who are active and doing what?

Some confusing / important terms

ASVs

amplicon sequencing

OTUs

DNA metabarcoding

metatranscriptomics

eDNA

rDNA phylotyping

metagenetics

metagenomics

microbiome analysis

MOTUs

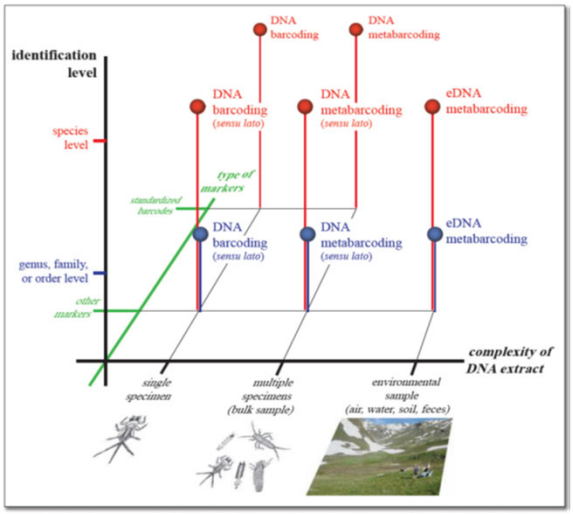
environmental sequencing

community profiling

marker gene analysis

Terminology

Pierre Taberlet



Taberlet *et al.* (2012) *Molecular Ecology*, **21**, 1789-1793.

DNA metabarcoding

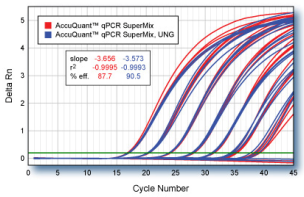
- Research questions:
  - Who are there?
  - Richness: How many taxa / operational taxonomic units (alpha/gamma diversity)?
  - Compositional differences (beta diversity)?
  - Which processes and drivers are shaping the communities?
  - Co-occurrence patterns → Interactions
- Qualitative data

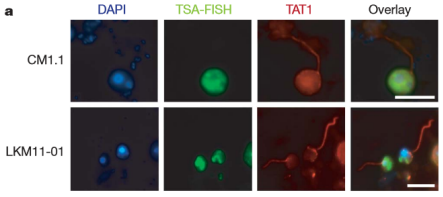
DNA metabarcoding

- Research questions:
  - Who are there?
  - Richness: How many taxa / operational taxonomic units (alpha/gamma diversity)?
  - Compositional differences (beta diversity)?
  - Which processes and drivers are shaping the communities?
  - Co-occurrence patterns → Interactions
- Qualitative data
- Quantitative data
  - Abundance: Who are common – who are rare…?

DNA metabarcoding

- Often functions as a first step looking into poorly characterized habitats and study systems
- Often generates hypotheses that can be addressed more in-depth with more taxon-specific approaches (e.g. qPCR or genomics) or genomics





DNA metabarcoding: From «wild west» towards an established approach

Primary phase
- Poor replication
- Lack of controls
- Lack of insight into important biases
- Poor bioinformatics approaches

EDITORIAL

MOLECULAR ECOLOGY WILEY

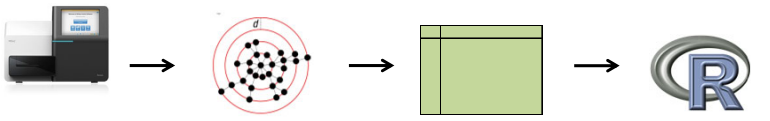
DNA metabarcoding—Need for robust experimental designs to draw sound ecological conclusions

Zinger et al. 2019, Molecular Ecology Resources

Established scientific approach with a set of widely accepted guidelines

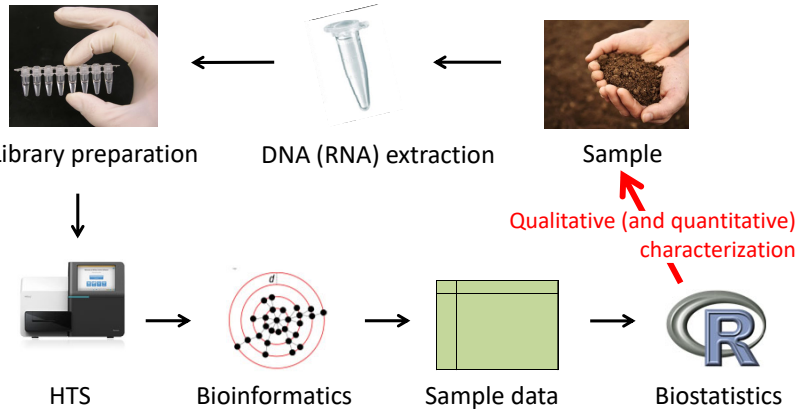
General workflow in DNA metabarcoding studies

Library preparation — DNA (RNA) extraction — Sample

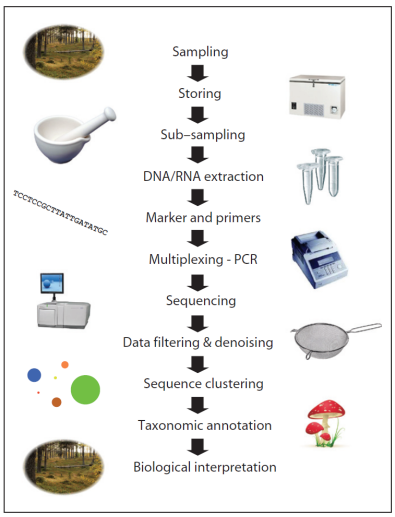


HTS — Bioinformatics — Sample data — Biostatistics

General workflow in DNA metabarcoding studies



DNA metabarcoding - many steps



Lindahl et al. 2013

Many steps…

… to go wrong

EDITORIAL

MOLECULAR ECOLOGY WILEY

DNA metabarcoding—Need for robust experimental designs to draw sound ecological conclusions

Zinger et al. 2019. Molecular Ecology Resources

RESEARCH ARTICLE

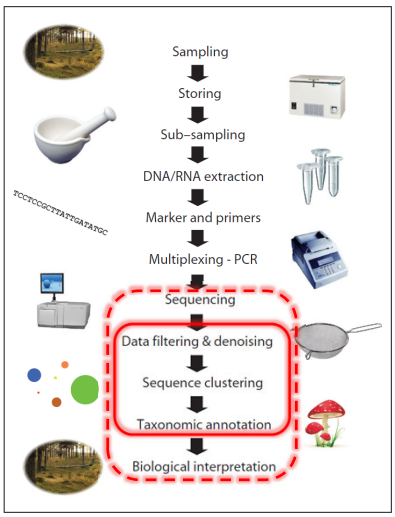
Methods in Ecology and Evolution

Scrutinizing key steps for reliable metabarcoding of environmental samples

Antton Alberdi[1] | Ostaizka Aizpurua[1] | M. Thomas P. Gilbert[1,2,3] | Kristine Bohmann[1,4]

Alberdi et al. 2017

DNA metabarcoding - many steps

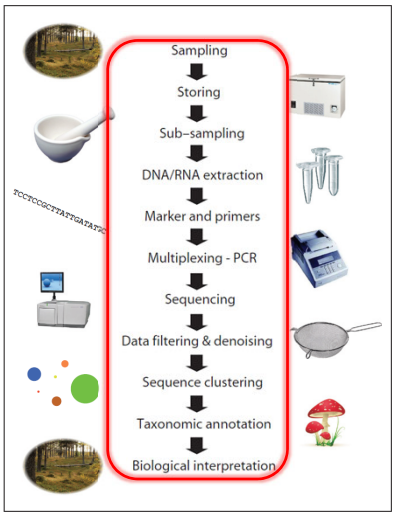


Lindahl et al. 2013

Many steps…

… to go wrong

DNA metabarcoding - many steps

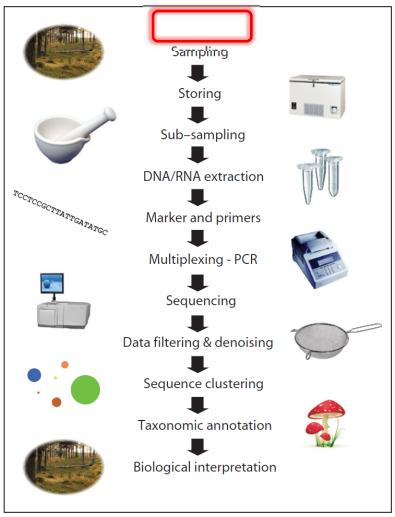


Lindahl et al. 2013

Many steps…

… to go wrong

DNA metabarcoding - many steps



Lindahl et al. 2013

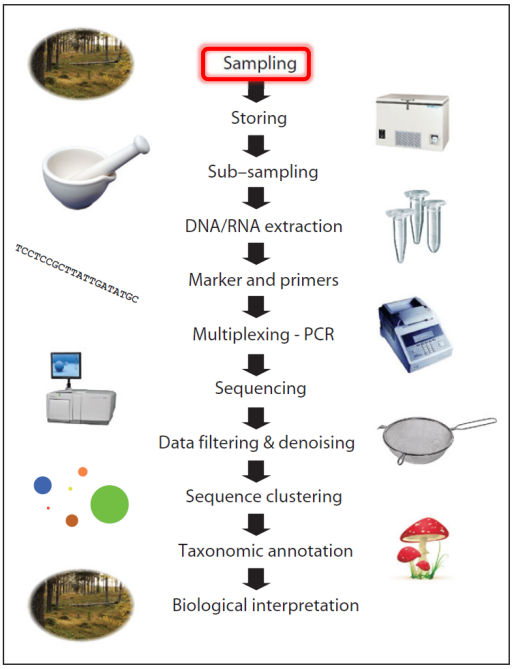
Many steps…

… to go wrong

If new study system – conduct a pilot?

- Which sampling scheme?
- How many replicates?
- Which extraction protocol
- Which primers?
- Which sequencing depth?
- Which sequencing technique?
- Etc.

→ Depends on the alpha, beta and gamma diversity, that you might not know anything about..



Representativeness (in space)

- Many communities highly heterogenious
- Should obtain samples that are representative
- They should not be too small
- If you are not interested in the small scale variation in itself → pool sub-samples?

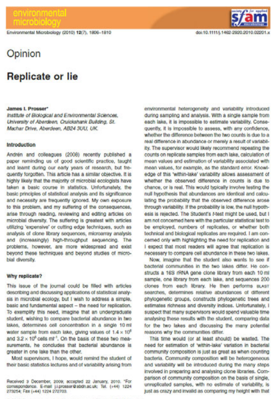
Fig. 5. What is the optimal relationship between primary sample size and the analytical sample volume (insert) and how can it come about? When sample size increases one can intuitively understand that the sample becomes more representative. But at the same time, today's analytical volumes continue to decrease (insert) as the analytical instruments become more and more precise. For all heterogeneous materials, there is consequently an intrinsic contradiction between primary sampling representativity and the instrumental analytical volume requirements. This is the root cause of all sampling and representativity issues.

Representativeness (in time)

- Many communities often display high temporal variation! Repeated temporal sampling?

Examples: Insects and fungal spores in the air

«Replicate or lie»



	Clone library analysis and pyrosequencing	
	Number of articles	% with replicates
Appl Environ Microbiol	60	23
Environ Microbiol	47	15
FEMS Microbiol Ecol	29	24
ISME J	23	13
Microbial Ecol	22	9
Total	181	18

It doesn't help that you are dealing with HTS data if you don't replicate properly!

Prosser JI. 2010, Environmental Microbiology

Received: 4 October 2017 | Revised: 10 May 2018 | Accepted: 14 May 2018
DOI: 10.1111/1755-0998.12907

INVITED TECHNICAL REVIEW

WILEY MOLECULAR ECOLOGY RESOURCES

Towards robust and repeatable sampling methods in eDNA-based studies

Ian A. Dickie[1,2] | Stephane Boyer[3,4] | Hannah L. Buckley[5] | Richard P. Duncan[6] | Paul P. Gardner[2] | Ian D. Hogg[7,8] | Robert J. Holdaway[9] | Gavin Lear[10] | Andreas Makiola[1] | Sergio E. Morales[11] | Jeff R. Powell[12] | Louise Weaver[13]

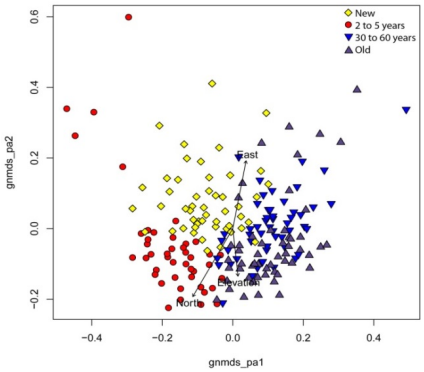
Abstract

DNA-based techniques are increasingly used for measuring the biodiversity (species presence, identity, abundance and community composition) of terrestrial and aquatic ecosystems. While there are numerous reviews of molecular methods and bioinformatic steps, there has been little consideration of the methods used to collect samples upon which these later steps are based. This represents a critical knowledge gap, as methodologically sound field sampling is the foundation for subsequent analyses. We reviewed field sampling methods used for metabarcoding studies of both terrestrial and freshwater ecosystem biodiversity over a nearly three-year period ($n = 75$). We found that 95% ($n = 71$) of these studies used subjective sampling methods and inappropriate field methods and/or failed to provide critical methodological information. It would be possible for researchers to replicate only 5% of the metabarcoding studies in our sample, a poorer level of reproducibility than for ecological studies in general. Our findings suggest greater attention to field sampling methods, and reporting is necessary in eDNA-based studies of biodiversity to ensure robust outcomes and future reproducibility. Methods must be fully and accurately reported, and protocols developed that minimize subjectivity. Standardization of sampling protocols would be one way to help to improve reproducibility and have additional benefits in allowing compilation and comparison of data from across studies.

Biological replicates

Fungal communities associated with mosses in different forest management types

Biological replicates are biologically distinct samples which show biological variation



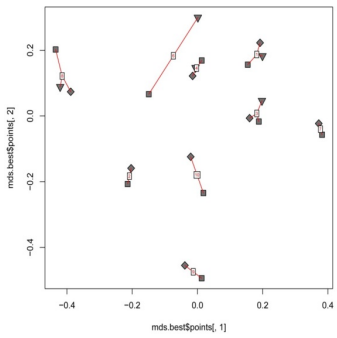
Davey *et al.* 2014. FEMS Microbial Ecology

Technical replicates

- Some samples should/can be analyzed multiple times
- Reveals the variability (experimental error) of the analysis → allows to set limits for what is meaningful and significant data
- How important?

Communities with low DNA content!!

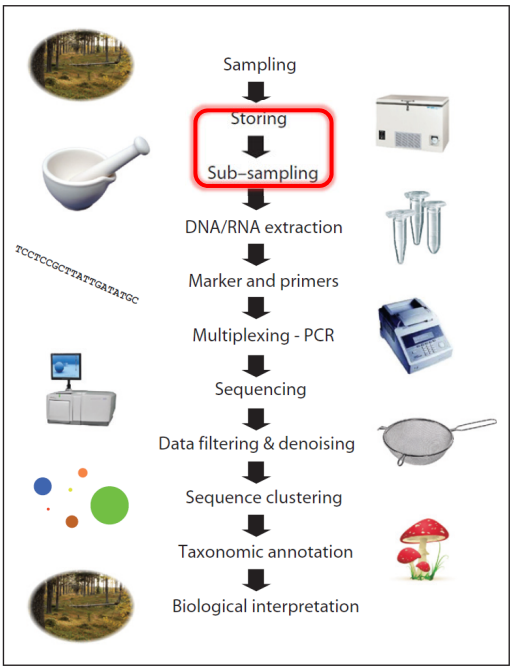
Technical replicates are repeated measurements of a sample, which show variation of the measuring equipment and protocol



Davey *et al.* 2014. FEMS Microbial Ecology

Sample types

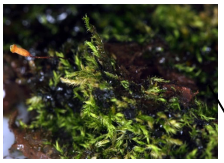
1. Biological replicates
2. Technical replicates



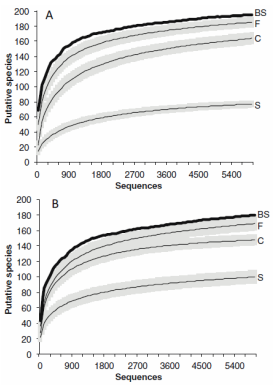
Storage

- Unnapporiate storage may introduce severe biases!
- Community members can respond quickly to altered conditions
- ‘Arrest’ the communities!
- Process the samples asap. If needed, long time storage at -80C often suggested

Storage



U’Ren *et al.* 2014. Tissue storage and primer selection influence pyrosequencing-based inferences of diversity and community composition of endolichenic and endophytic fungi. Mol Ecol Res.



F = Fresh
C = CTAB
S = Silica

Sub-sampling and homogenization

- Protocols for nucleic acid extraction are normally based on small amounts
- Field samples are often much larger, and careful homogenization of material is required to reduce the sample to a smaller but still representative subsamples
- The most commonly used techniques are bead beating and/or crushing in liquid nitrogen.

Homogenization of samples

- After homogenization, new spatial structures may easily be created in the samples, for example by density fractionation at the slightest bumping.
- To make proper comparisons: DNA should be extracted from equivalent amounts of starting material!

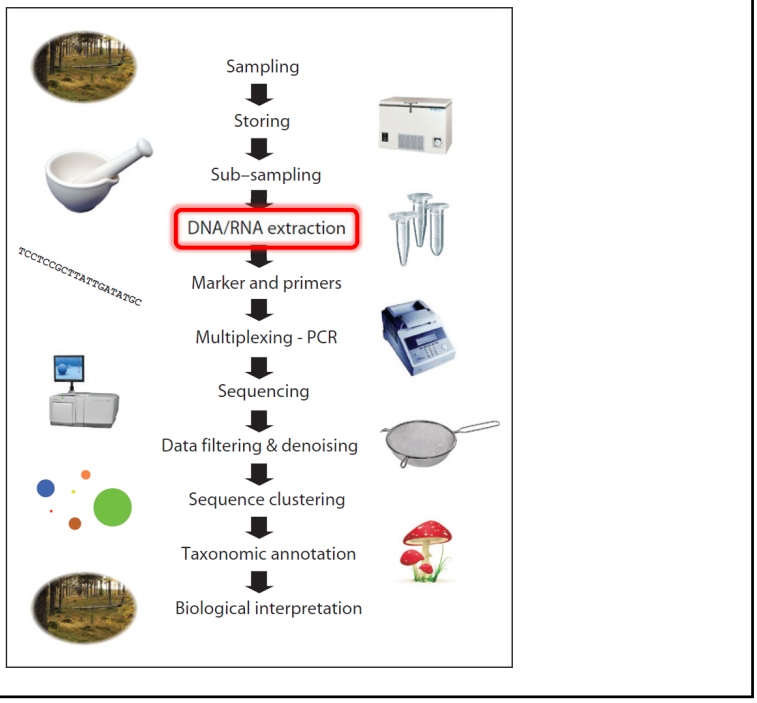
MOLECULAR ECOLOGY

Molecular Ecology (2012) 21, 1816–1820 doi: 10.1111/j.1365-294X.2011.05317.x

Soil sampling and isolation of extracellular DNA from large amount of starting material suitable for metabarcoding studies

PIERRE TABERLET, SOPHIE M. PRUD'HOMME, ETIENNE CAMPIONE, JULIEN ROY, CHRISTIAN MIQUEL, WASIM SHEHZAD, LUDOVIC GIELLY, DELPHINE RIOUX, PHILIPPE CHOLER, JEAN-CHRISTOPHE CLÉMENT, CHRISTELLE MELODELIMA, FRANÇOIS POMPANON and ERIC COISSAC

Laboratoire d'Ecologie Alpine, CNRS UMR 5553, Université Joseph Fourier, BP 53, F-38041 Grenoble Cedex 9, France



DNA extraction

An evaluation of commercial DNA extraction kits for the isolation of bacterial spore DNA from soil

S.M. Dineen[1,2], R. Aranda IV[1,2], D.L. Anders[3] and J.M. Robertson[2]

1 Visiting Scientist, Federal Bureau of Investigation Laboratory, Quantico, VA, USA
2 Counterterrorism and Forensic Science Research Unit, Federal Bureau of Investigation Laboratory, Quantico, VA, USA
3 Hazardous Materials Science Response Unit, Federal Bureau of Investigation Laboratory, Quantico, VA, USA

NOTE / NOTE

Influence of DNA extraction and PCR amplification on studies of soil fungal communities based on amplicon sequencing

Lihui Xu, Sabine Ravnskov, John Larsen, and Jørgens Nicolaisen

Molecular biology, genetics and biotechnology

Effect of DNA extraction and sample preservation method on rumen bacterial population

Katerina Fliegerova[a,*], Ilma Tapio[b], Aurelie Bonin[c], Jakub Mrazek[a], Maria Luisa Callegari[c], Paolo Bani[c], Alireza Bayat[d], Johanna Vilkki[b], Jan Kopecný[a], Kevin J. Shingfi[a,d], Frederic Boyer[c], Eric Coissac[c], Pierre Taberlet[c], R. John Wallace[f]

PLOS

Effect of DNA Extraction Methods and Sampling Techniques on the Apparent Structure of Cow and Sheep Rumen Microbial Communities

Gemma Henderson[1], Faith Cox[1], Sandra Kittelmann[1], Vahideh Heidarian Miri[1], Michael Zethof[1], Samantha J. Noel[1,2], Garry C. Waghorn[3], Peter H. Janssen[1]

The Impact of Different DNA Extraction Kits and Laboratories upon the Assessment of Human Gut Microbiota Composition by 16S rRNA Gene Sequencing

Nicholas A. Kennedy[1], Alan W. Walker[2], Susan H. Berry[3], Sylvia H. Duncan[4], Freda M. Farquarson[4], Petra Louis[4], John M. Thomson[5], UK IBD Genetics Consortium[†], Jack Satsangi[1], Harry J. Flint[4], Julian Parkhill[2], Charlie W. Lees[1]*, Georgina L. Hold[3]*

DNA extraction

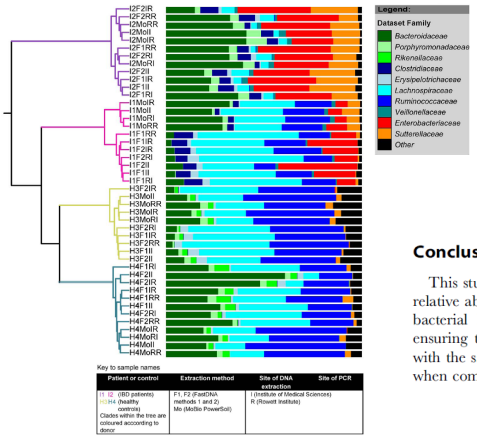
- Should yield high and uniform amounts of DNA
- Concentration of PCR inhibitors minimized
- Same protocol for all samples!
- If no proper literature are available on your study system → conduct a pilot?!

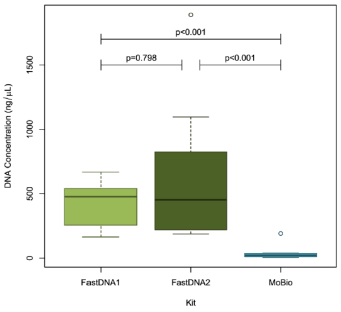
- MoBio Power Soil?
- FastDNA kit for Soil?
- EZNA Soil kit?
- CTAB + cleanup kit?

DNA extraction

The Impact of Different DNA Extraction Kits and Laboratories upon the Assessment of Human Gut Microbiota Composition by 16S rRNA Gene Sequencing

Nicholas A. Kennedy[1], Alan W. Walker[2], Susan H. Berry[3], Sylvia H. Duncan[4], Freda M. Farquarson[4], Petra Louis[4], John M. Thomson[5], UK IBD Genetics Consortium*, Jack Satsangi[1], Harry J. Flint[4], Julian Parkhill[2], Charlie W. Lees[1]*, Georgina L. Hold[3]*





Conclusions

This study demonstrates important differences in the yield and relative abundance of key bacterial families for kits used to isolate bacterial DNA from stool. This highlights the importance of ensuring that all samples to be analyzed together are prepared with the same DNA extraction method, and the need for caution when comparing studies that have used different methods.

Note: Lack of true replicates

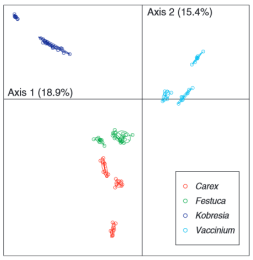
DNA extraction

- Both extracellular and intracelluar DNA are normally co-extracted.
- Method for extraction of extracellular DNA from large amount of starting material:

MOLECULAR ECOLOGY
Molecular Ecology (2012) 21, 1816–1820 doi: 10.1111/j.1365-294X.201

Soil sampling and isolation of extracellular DNA from large amount of starting material suitable for metabarcoding studies

PIERRE TABERLET, SOPHIE M. PRUD'HOMME, ETIENNE CAMPIONE, JULIEN ROY, CHRISTIAN MIQUEL, WASIM SHEHZAD, LUDOVIC GIELLY, DELPHINE RIOUX, PHILIPPE CHOLER, JEAN-CHRISTOPHE CLÉMENT, CHRISTELLE MELODELIMA, FRANÇOIS POMPANON and ERIC COISSAC
Laboratoire d'Ecologie Alpine, CNRS UMR 5553, Université Joseph Fourier, BP 53, F-38041 Grenoble Cedex 9, France



Replicates!

DNA extraction

- Should yield high and uniform amounts of DNA
- Concentration of PCR inhibitors minimized
- Same protocol for all samples!
- If no proper literature are available on your study system → conduct a pilot?!
- Extraction negatives!

- MoBio Power Soil?
- FastDNA kit for Soil?
- EZNA Soil kit?
- CTAB + cleanup kit?

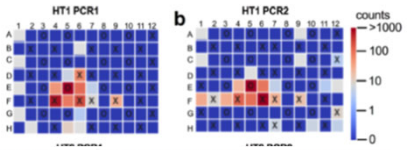
DNA extraction

mSystems®

RESEARCH ARTICLE
Novel Systems Biology Techniques

Quantifying and Understanding Well-to-Well Contamination in Microbiome Research

Jeremiah J. Minich,[a] Jon G. Sanders,[b] Amnon Amir,[b] Greg Humphrey,[b] Jack A. Gilbert,[c,d] Rob Knight[b,e,f]

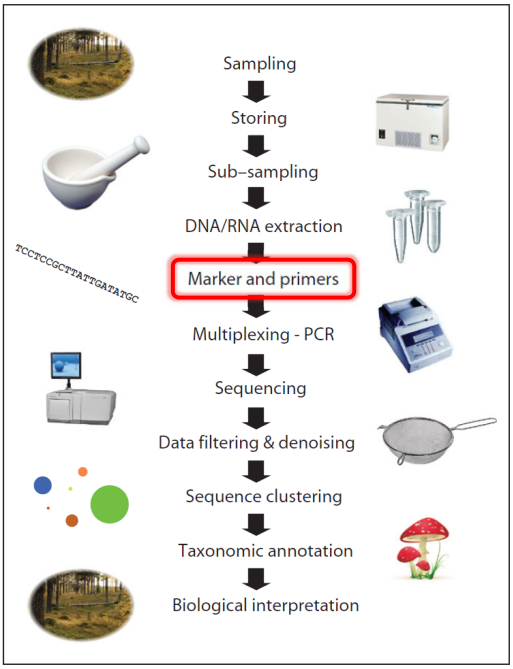


ABSTRACT Microbial sequences inferred as belonging to one sample may not have originated from that sample. Such contamination may arise from laboratory or reagent sources or from physical exchange between samples. This study seeks to rigorously assess the behavior of this often-neglected between-sample contamination. Using unique bacteria, each assigned a particular well in a plate, we assess the frequency at which sequences from each source appear in other wells. We evaluate the effects of different DNA extraction methods performed in two laboratories using a consistent plate layout, including blanks and low-biomass and high-biomass samples. Well-to-well contamination occurred primarily during DNA extraction and, to a lesser extent, in library preparation, while barcode leakage was negligible. Laboratories differed in the levels of contamination. Extraction methods differed in their occurrences and levels of well-to-well contamination, with plate methods having more well-to-well contamination and single-tube methods having higher levels of background contaminants. Well-to-well contamination occurred primarily in neighboring samples, with rare events up to 10 wells apart. This effect was greatest in samples with lower biomass and negatively impacted metrics of alpha and beta diversity. Our work emphasizes that sample contamination is a combination of cross talk from nearby wells and background contaminants. To reduce well-to-well effects, samples should be randomized across plates, samples of similar biomasses should be processed together, and manual single-tube extractions or hybrid plate-based cleanups should be employed. Researchers should avoid simplistic removals of taxa or operational taxonomic units (OTUs) appearing in negative controls, as many will be microbes from other samples rather than reagent contaminants.

Minich et al. 2019, mSystems

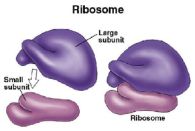
Sample types

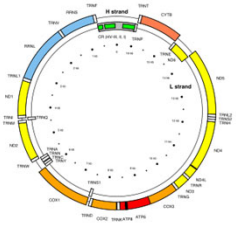
1. Biological replicates
2. Technical replicates
3. Extraction negatives



Markers used in DNA metabarcoding

- Standard markers (<500 bp):
  - 18S: Eukaryotes
  - 16S: Bacteria/archaea
  - ITS: Fungi & plants
  - COI: Metazoa
  - *Rbc*L: Plants
  - *trn*L: Plants





Markers in DNA metabarcoding

- The ideal marker should:
  - Have primer sites that are shared by all target organisms
  - Be easy to amplify
  - Be of appropriate length for efficient amplification and sequencing
  - Be of similar length
  - No intragenomic variation (i.e. no paralogs)
  - Similar number of copies
  - Be possible to align
  - Have high interspecific variation
  - Have low intraspecific variation
- No known markers meet all these requirements!

Markers in DNA metabarcoding

- The ideal marker should:
  - Have primer sites that are shared by all target organisms
  - Be easy to amplify
  - Be of appropriate length for efficient amplification and sequencing
  - Be of similar length
  - No intragenomic variation (i.e. no paralogs)
  - Similar number of copies
  - Be possible to align
  - Have high interspecific variation
  - Have low intraspecific variation
- No known markers meet all these requirements!

ITS and primer bias

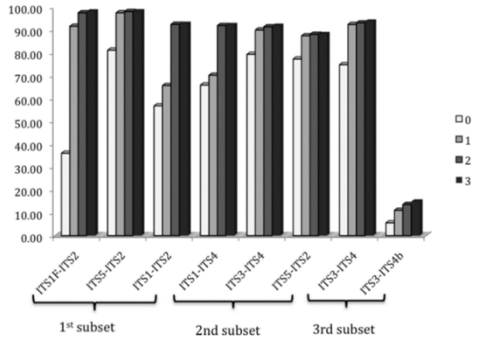
Bellemain et al. BMC Microbiology 2010, 10:189
http://www.biomedcentral.com/1471-2180/10/189

BMC Microbiology

RESEARCH ARTICLE Open Access

ITS as an environmental DNA barcode for fungi: an *in silico* approach reveals potential PCR biases

Eva Bellemain*[1], Tor Carlsen[2], Christian Brochmann[1], Eric Coissac[3], Pierre Taberlet[3] and Håvard Kauserud[2]



Bellemain et al. 2010, BMC Microbiology

16S and primer bias

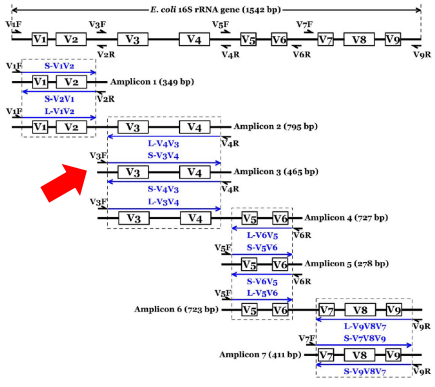
OPEN ACCESS Freely available online PLOS ONE

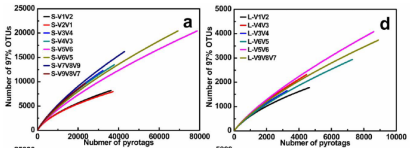
Biased Diversity Metrics Revealed by Bacterial 16S Pyrotags Derived from Different Primer Sets

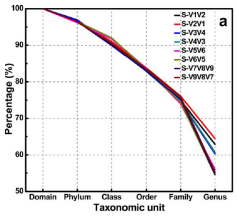
Lin Cai[1], Lin Ye[1], Amy Hin Yan Tong[2], Si Lok[2], Tong Zhang[1]*

1 Environmental Biotechnology Laboratory, Department of Civil Engineering, The University of Hong Kong, Hong Kong SAR, China, 2 Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong SAR, China

Bacterial communities in sludges







16S and primer bias

OPEN ACCESS Freely available online

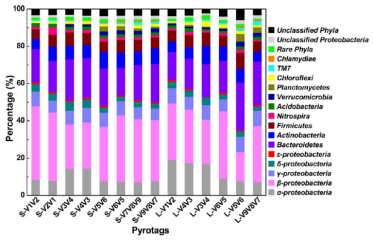
PLOS ONE

Biased Diversity Metrics Revealed by Bacterial 16S Pyrotags Derived from Different Primer Sets

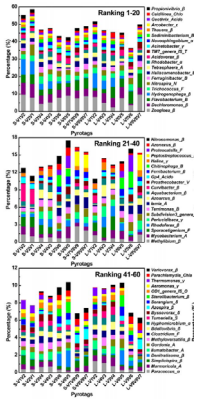
Lin Cai[1], Lin Ye[1], Amy Hin Yan Tong[2], Si Lok[2], Tong Zhang[1]*

1 Environmental Biotechnology Laboratory, Department of Civil Engineering, The University of Hong Kong, Hong Kong SAR, China, 2 Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong SAR, China

Phylum/class level



Genus level (top 60)



The 18S marker

OPEN ACCESS Freely available online

PLOS ONE

Characterization of the 18S rRNA Gene for Designing Universal Eukaryote Specific Primers

Kenan Hadziavdic[1], Katrine Lekang[1], Anders Lanzen[1,2,3], Inge Jonassen[2,4], Eric M. Thompson[1,5], Christofer Troedsson[6]*

1 Department of Biology, University of Bergen, Bergen, Norway, 2 Uni Computing, Uni Research AS, Bergen, Norway, 3 Department of Ecology and Natural Resources, NEIKER-Tecnalia, Derio, Spain, 4 Department of Informatics, University of Bergen, Bergen, Norway, 5 Sars International Centre for Marine Molecular Biology, University of Bergen, Bergen, Norway, 6 Uni Environment, Uni Research AS, Bergen, Norway

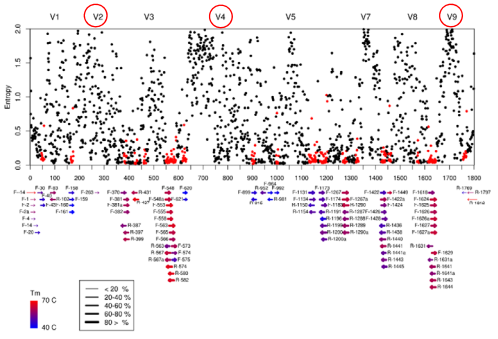


Figure 1. Nucleotide variability in 18S rRNA genes. Shannon entropy values of all eukaryotic alignment positions from the SILVA database along the 18S rRNA gene of *Saccharomyces cerevisiae*. Red dots mark consecutive nucleotide positions where at least 90% of ≥10 nt have entropy values lower than 0.2. The highly variable regions of 18S rRNA gene are denoted V1 to V9. In total, 100 primers targeting eukaryotes from the literature as well as those designed in this study are positioned along the reference sequence. The direction of the arrows indicates the orientation of the primers. The color denotes the melting temperature, and the thickness of the arrows represents the eukaryote universality of the primers. doi:10.1371/journal.pone.0087624.g001

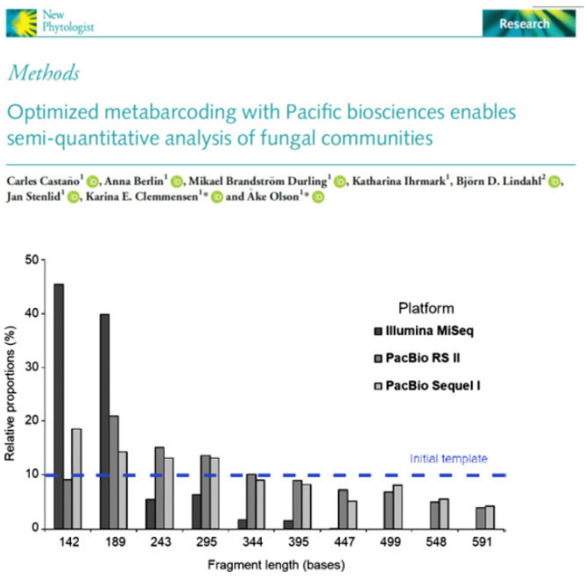
A characterization of the variable and conserved regions within the 18S rRNA gene Using the SILVA database with > 500,000 SSU rRNA sequences

Best primer pair: 566 and 1200. When allowing one mismatch, almost all phyla covered.

Markers in DNA metabarcoding

- The ideal marker should:
  - Have primer sites that are shared by all target organisms
  - Be easy to amplify
  - Be of appropriate length for efficient amplification and sequencing
  - Be of similar length
  - No intragenomic variation (i.e. no paralogs)
  - Be possible to align
  - Have high interspecific variation
  - Have low intraspecific variation
- No known markers meet all these requirements!

With length variation → lenght biases intrudced during both PCR and sequencing!



Markers in DNA metabarcoding

- The ideal marker should:
  - Have primer sites that are shared by all target organisms
  - Be easy to amplify
  - Be of appropriate length for efficient amplification and sequencing
  - Be of similar length
  - No intragenomic variation (i.e. no paralogs)
  - Similar number of copies
  - Be possible to align
  - Have high interspecific variation
  - Have low intraspecific variation
- No known markers meet all these requirements!

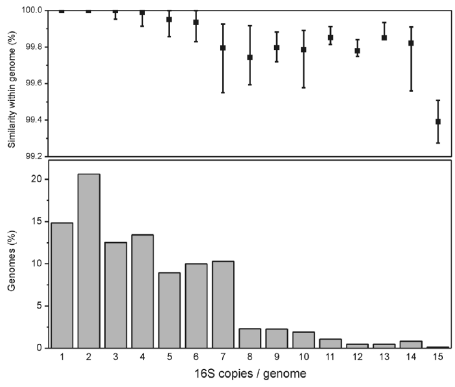
(Intra)genomic variability in 16S

OPEN ACCESS Freely available online PLOS ONE

The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses

Tomáš Větrovský, Petr Baldrian*

Laboratory of Environmental Microbiology, Institute of Microbiology of the Academy of Sciences of the Czech Republic, Praha, Czech Republic



1,690 bacterial genomes with 7,081 16S copies analyzed

The same for 18S and eukaryotes!

Markers in DNA metabarcoding

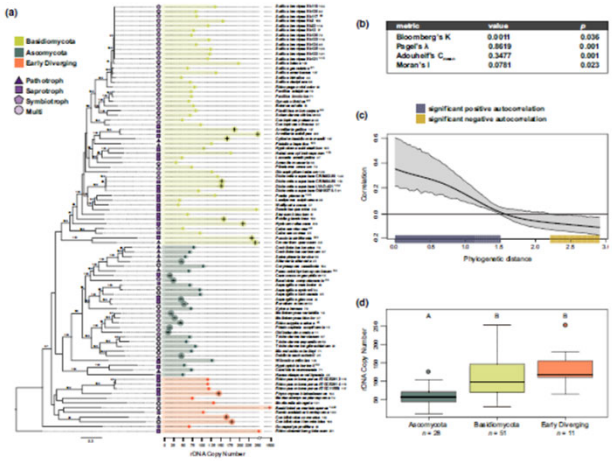
- The ideal marker should:
  - Have primer sites that are shared by all target organisms
  - Be easy to amplify
  - Be of appropriate length for efficient amplification and sequencing
  - Be of similar length
  - No intragenomic variation (i.e. no paralogs)
  - Similar number of copies
  - Be possible to align
  - Have high interspecific variation
  - Have low intraspecific variation
- No known markers meet all these requirements!

ORIGINAL ARTICLE

WILEY MOLECULAR ECOLOGY

Genome-based estimates of fungal rDNA copy number variation across phylogenetic scales and ecological lifestyles

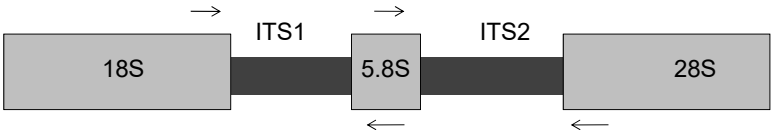
Lotus A. Lofgren[1] | Jessie K. Uehling[2] | Sara Branco[3] | Thomas D. Bruns[2] | Francis Martin[4] | Peter G. Kennedy[1,5]

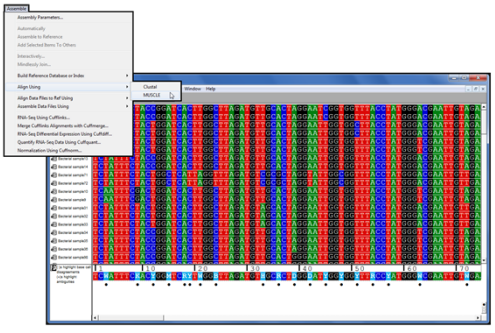


Markers in DNA metabarcoding

- The ideal marker should:
  - Have primer sites that are shared by all target organisms
  - Be easy to amplify
  - Be of appropriate length for efficient amplification and sequencing
  - Be of similar length
  - No intragenomic variation (i.e. no paralogs)
  - Similar number of copies
  - Be possible to align (in a multiple alignment)
  - Have high interspecific variation
  - Have low intraspecific variation
- No known markers meet all these requirements!

Possible to align

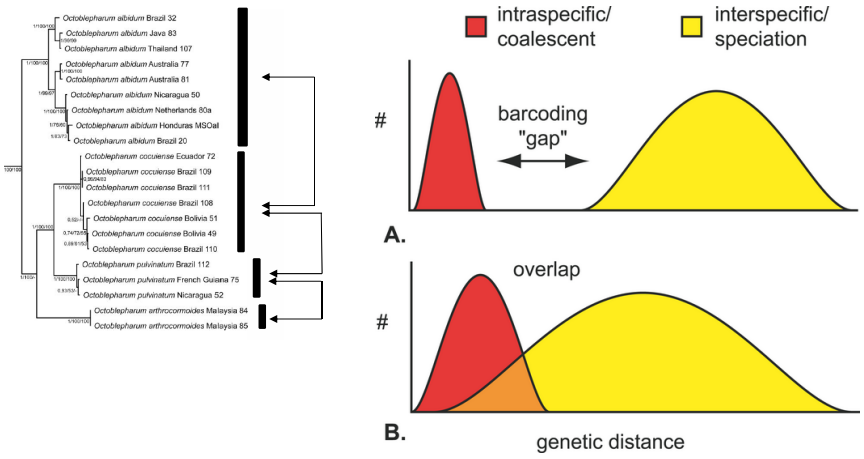




Markers in DNA metabarcoding

- The ideal marker should:
  - Have primer sites that are shared by all target organisms
  - Be easy to amplify
  - Be of appropriate length for efficient amplification and sequencing
  - Be of similar length
  - No intragenomic variation (i.e. no paralogs)
  - Be possible to align
  - Possess high interspecific variation
  - Possess low intraspecific variation
- No known markers meet all these requirements!

The barcoding gap



How conserved/variable are the marker?

- 18S and 16S: Low variability, low intraspecific variation, low interspecific variation
- ITS: High variability, high intraspecific variation, high 'interspecific' variation

- Impact how the bioinformatics analyses should be conducted (i.e. no single way)!!

Multiple markers/primers?

Drummond *et al. GigaScience* (2015) 4:46
DOI 10.1186/s13742-015-0086-1

(GIGA)n SCIENCE

RESEARCH Open Access

Evaluating a multigene environmental DNA approach for biodiversity assessment

Alexei J. Drummond[1,2*], Richard D. Newcomb[1,3,4], Thomas R. Buckley[1,3,5], Dong Xie[1,2], Andrew Dopheide[1,3,4], Benjamin CM Potter[1,3], Joseph Heled[1,2], Howard A. Ross[1,3], Leah Tooman[1,4], Stefanie Grosser[1,5], Duckchul Park[5], Nicholas J. Demetras[8], Mark I. Stevens[6,7], James C. Russell[1,3,9], Sandra H. Anderson[3], Anna Carter[1,10] and Nicola Nelson[1,10]

- Inherent problem in compex samples: Difficult to link alleles/variants across DNA markers/loci
- Often lack proper reference databases for multiple markers.

Long-read metabarcoding

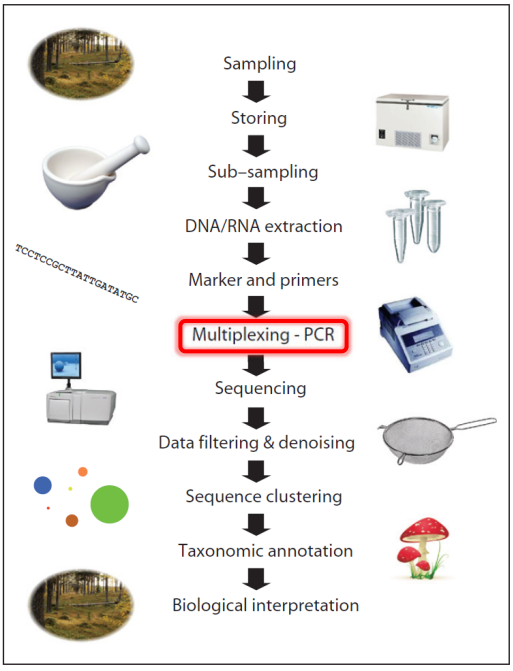
RESOURCE ARTICLE

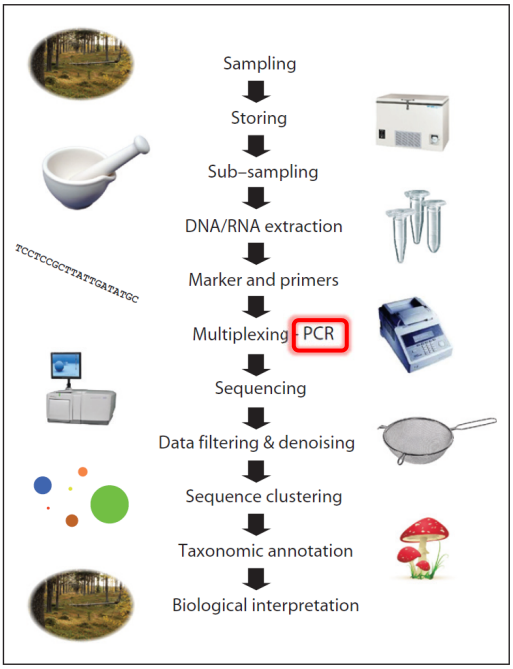
MOLECULAR ECOLOGY RESOURCES WILEY

Long-read metabarcoding of the eukaryotic rDNA operon to phylogenetically and taxonomically resolve environmental diversity

Mahwash Jamy[1] | Rachel Foster[2] | Pierre Barbera[3] | Lucas Czech[3] | Alexey Kozlov[3] | Alexandros Stamatakis[3,4] | Gary Bending[5] | Sally Hilton[5] | David Bass[2,6] | Fabien Burki[1]

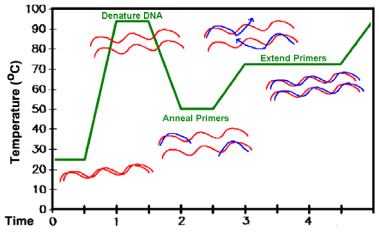
- Phylogenetic framework → more secure taxonomic placement
- Comes with extra challenges
  - Harder to amplify
  - More chimeric sequences
  - Lower sequencing depth (PacBio or Nanopore)





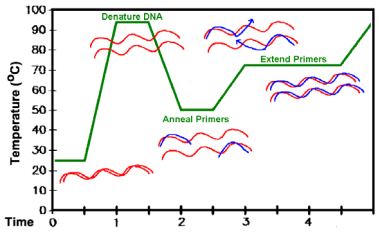
PCR

- Different relevant factors during PCR:
  - Which polymerase enzyme (proofreading or not)?
  - Which RAMP speed?
  - How many cycles?
  - Which annealing temperature?
  - Multiple/replicate PCR reactions?
  - PCR negatives!



PCR

- Different relevant factors during PCR:
  - Which polymerase enzyme (proofreading or not)?
  - Which RAMP speed?
  - How many cycles?
  - Which annealing temperature?
  - Multiple/replicate PCR reactions?
  - PCR negatives!



Polymerase enzyme

The Journal of Microbiology (2012) Vol. 50, No. 6, pp. 1071–1074
DOI 10.1007/s12275-012-2642-z

NOTE

Effects of PCR Cycle Number and DNA Polymerase Type on the 16S rRNA Gene Pyrosequencing Analysis of Bacterial Communities§

Jae-Hyung Ahn, Byung-Yong Kim, Jaekyeong Song, and Hang-Yeon Weon*

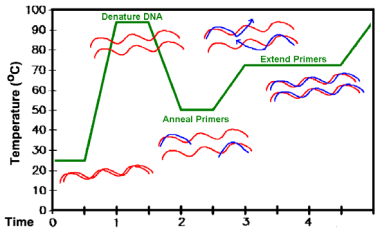
Agricultural Microbiology Division, National Academy of Agricultural Science, Rural Development Administration, Suwon 441-707, Republic of Korea

(Received November 19, 2012 / Accepted December 11, 2012)

. The bacterial richness was overestimated at increased PCR cycle number mostly due to the occurence of chimeric sequences, and this was more serious with a DNA polymerase having proofreading activity than with *Taq* DNA polymerase. These results suggest that PCR cycle number must be kept as low as possible for accurate estimation of bacterial richness and that particular care must be taken when a DNA polymerase having proofreading activity is used.

PCR

- Different relevant factors during PCR:
  - Which polymerase enzyme (proofreading or not)?
  - Which RAMP speed?
  - How many cycles?
  - Which annealing temperature?
  - Multiple/replicate PCR reactions?
  - PCR negatives!



Polymerase enzyme

The Journal of Microbiology (2012) Vol. 50, No. 6, pp. 1071–1074
DOI 10.1007/s12275-012-2642-z

NOTE

Effects of PCR Cycle Number and DNA Polymerase Type on the 16S rRNA Gene Pyrosequencing Analysis of Bacterial Communities§

Jae-Hyung Ahn, Byung-Yong Kim, Jaekyeong Song, and Hang-Yeon Weon*

Agricultural Microbiology Division, National Academy of Agricultural Science, Rural Development Administration, Suwon 441-707, Republic of Korea

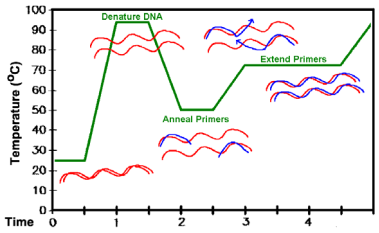
(Received November 19, 2012 / Accepted December 11, 2012)

. The bacterial richness was overestimated at increased PCR cycle number mostly due to the occurence of chimeric sequences, and this was more serious with a DNA polymerase having proofreading activity than with *Taq* DNA polymerase. These results suggest that PCR cycle number must be kept as low as possible for accurate estimation of bacterial richness and that particular care must be taken when a DNA polymerase having proofreading activity is used.

Keep the number of cycles as low as possible!

PCR

- Different relevant factors during PCR:
  - Which polymerase enzyme (proofreading or not)?
  - Which RAMP speed?
  - How many cycles?
  - Which annealing temperature?
  - Multiple/replicate PCR reactions?
  - PCR negatives!

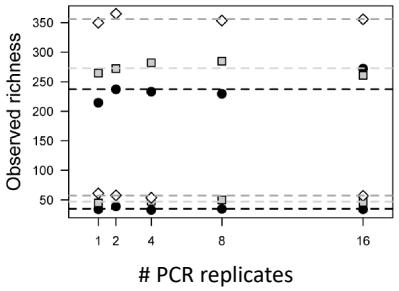


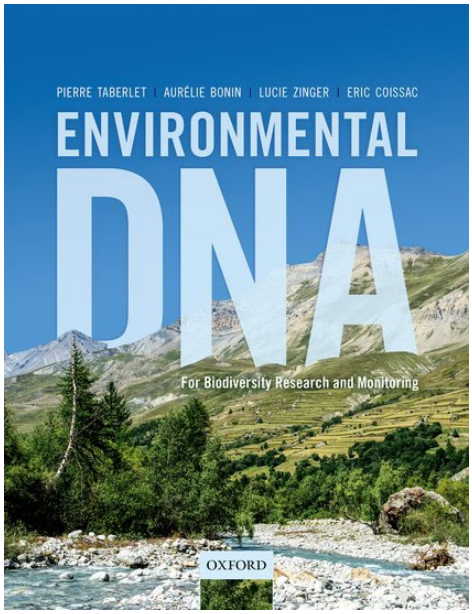
PCR replicates?

OPEN ACCESS Freely available online
PLOS ONE

Sequence Depth, Not PCR Replication, Improves Ecological Inference from Next Generation DNA Sequencing

Dylan P. Smith, Kabir G. Peay*
Department of Biology, Stanford University, Stanford, California, United States of America





Taberlet et al: Multiple PCR amplifications and remove outliers.

Probably very context dependent: Low DNA concentrations means that stochastisity during PCR is higher → The need for multiple amplicons is higher

Sample types

1. Biological replicates
2. Technical replicates
3. Extraction negatives

+ PCR replicates: A type of technical replicates

PCR

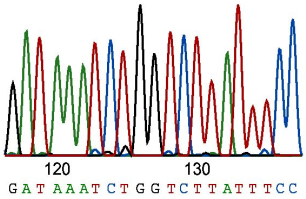
- Different relevant factors during PCR:
  - Which polymerase enzyme (proofreading or not)?
  - Which RAMP speed?
  - How many cycles?
  - Which annealing temperature?
  - Multiple/replicate PCR reactions?
  - PCR negatives!

Sample types

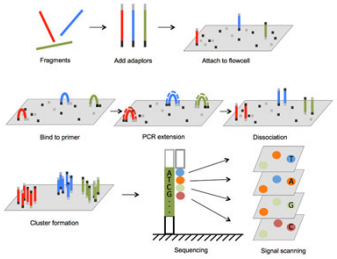
1. Biological replicates
2. Technical replicates
3. Extraction negatives
4. PCR negatives

PCR-induced errors

- PCR mutations: polymerase enzymes introduce erroneous nucleotides now and then, even those enzymes with proof-reading activity
  - Dependent on the technology whether these becomes «visible»
  - In classic (direct) Sanger sequencing such errors become «diluted»
  - In methods where your final sequences are derived from one single DNA template, they become visible and must be corrected for!

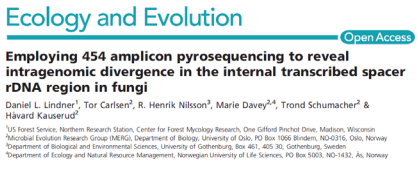


Sanger



Illumina

PCR-induced errors



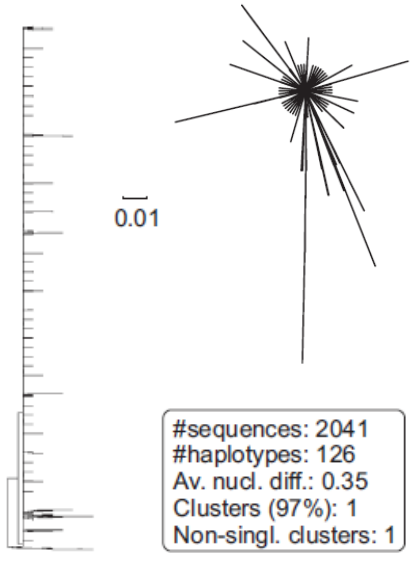
Ecology and Evolution

Open Access

Employing 454 amplicon pyrosequencing to reveal intragenomic divergence in the internal transcribed spacer rDNA region in fungi

Daniel L. Lindner[1], Tor Carlsen[2], R. Henrik Nilsson[3], Marie Davey[2,4], Trond Schumacher[2] & Håvard Kauserud[2]

[1]US Forest Service, Northern Research Station, Center for Forest Mycology Research, One Gifford Pinchot Drive, Madison, Wisconsin
[2]Microbial Evolution Research Group (MERG), Department of Biology, University of Oslo, PO Box 1066 Blindern, NO-0316, Oslo, Norway
[3]Department of Biological and Environmental Sciences, University of Gothenburg, Box 461, 405 30, Gothenburg, Sweden
[4]Department of Ecology and Natural Resource Management, Norwegian University of Life Sciences, PO Box 5003, NO-1432, Ås, Norway



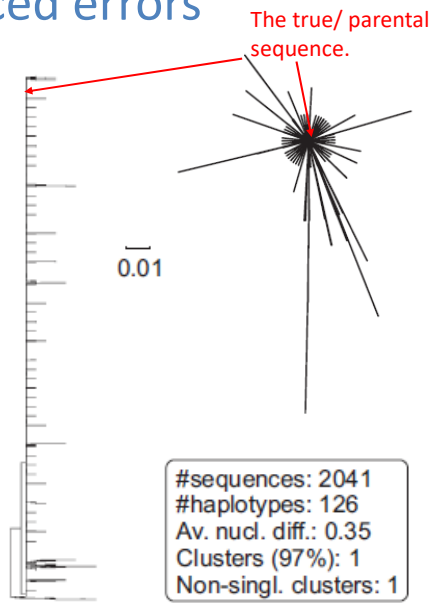
PCR-induced errors

Ecology and Evolution
Open Access

Employing 454 amplicon pyrosequencing to reveal intragenomic divergence in the internal transcribed spacer rDNA region in fungi

Daniel L. Lindner[1], Tor Carlsen[2], R. Henrik Nilsson[3], Marie Davey[2,4], Trond Schumacher[2] & Håvard Kauserud[2]

[1]US Forest Service, Northern Research Station, Center for Forest Mycology Research, One Gifford Pinchot Drive, Madison, Wisconsin
[2]Microbial Evolution Research Group (MERG), Department of Biology, University of Oslo, PO Box 1066 Blindern, NO-0316, Oslo, Norway
[3]Department of Biological and Environmental Sciences, University of Gothenburg, Box 461, 405 30, Gothenburg, Sweden
[4]Department of Ecology and Natural Resource Management, Norwegian University of Life Sciences, PO Box 5003, NO-1432, Ås, Norway



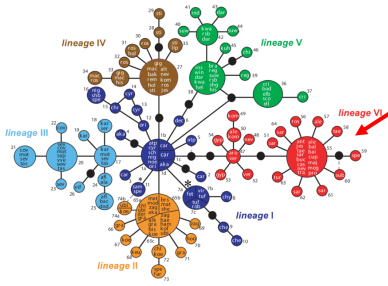
PCR-induced errors

Ecology and Evolution
Open Access

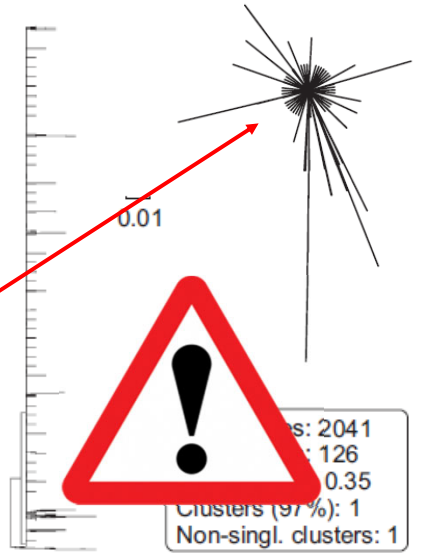
Employing 454 amplicon pyrosequencing to reveal intragenomic divergence in the internal transcribed spacer rDNA region in fungi

Daniel L. Lindner[1], Tor Carlsen[2], R. Henrik Nilsson[3], Marie Davey[2,4], Trond Schumacher[2] & Håvard Kauserud[2]

[1]US Forest Service, Northern Research Station, Center for Forest Mycology Research, One Gifford Pinchot Drive, Madison, Wisconsin
[2]Microbial Evolution Research Group (MERG), Department of Biology, University of Oslo, PO Box 1066 Blindern, NO-0316, Oslo, Norway
[3]Department of Biological and Environmental Sciences, University of Gothenburg, Box 461, 405 30, Gothenburg, Sweden
[4]Department of Ecology and Natural Resource Management, Norwegian University of Life Sciences, PO Box 5003, NO-1432, Ås, Norway

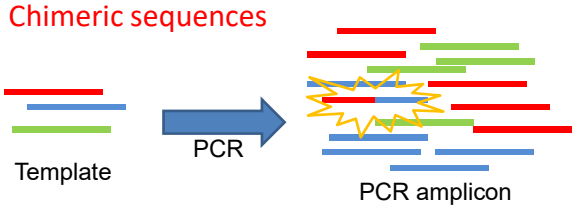


Same structure as natural variaton…



PCR-induced errors

- Chimeric sequences



MOLECULAR ECOLOGY RESOURCES
Molecular Ecology Resources (2016) doi: 10.1111/1755-0998.12622

ITS all right mama: investigating the formation of chimeric sequences in the ITS2 region by DNA metabarcoding analyses of fungal mock communities of different complexities

ANDERS BJØRNSGAARD AAS, MARIE LOUISE DAVEY and HÅVARD KAUSERUD
Section for Genetics and Evolutionary Biology (Evogene), Department of Biosciences, University of Oslo, P.O. Box 1066 Blindern, NO-0316 Oslo, Norway

The level of chimeric sequences depends on how variable the marker is!

Can reduce the problem with certain PCR settings, including long extension time and low number of cycles.

Polymerase enzyme

The Journal of Microbiology (2012) Vol. 50, No. 6, pp. 1071–1074 DOI 10.1007/s12275-012-2642-z
Copyright © 2012, The Microbiological Society of Korea

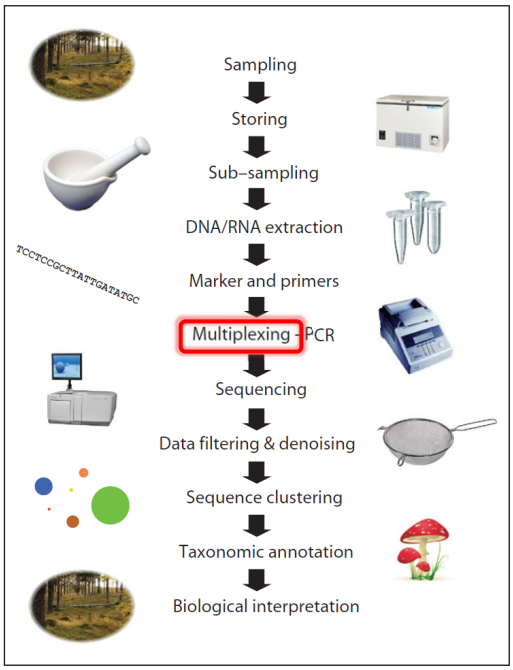
NOTE

Effects of PCR Cycle Number and DNA Polymerase Type on the 16S rRNA Gene Pyrosequencing Analysis of Bacterial Communities§

Jae-Hyung Ahn, Byung-Yong Kim, Jaekyeong Song, and Hang-Yeon Weon*
Agricultural Microbiology Division, National Academy of Agricultural Science, Rural Development Administration, Suwon 441-707, Republic of Korea

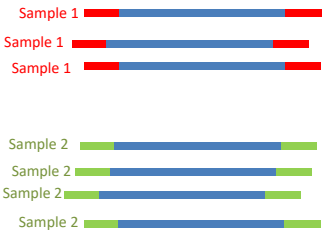
(Received November 19, 2012 / Accepted December 11, 2012)

. The bacterial richness was overestimated at increased PCR cycle number mostly due to the occurence of chimeric sequences, and this was more serious with a DNA polymerase having proofreading activity than with *Taq* DNA polymerase. These results suggest that PCR cycle number must be kept as low as possible for accurate estimation of bacterial richness and that particular care must be taken when a DNA polymerase having proofreading activity is used.



Different approaches to index the samples:

- Directly during the PCR
- Nested PCR approach



Tag switching

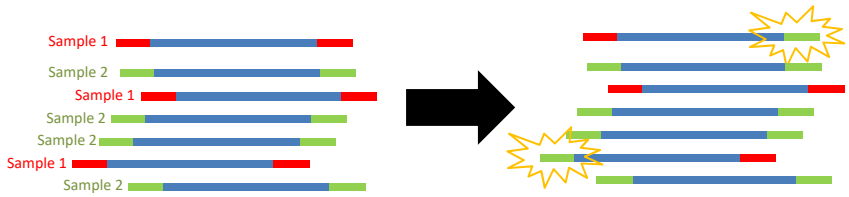
Commentary

Don't make a mista(g)ke: is tag switching an overlooked source of error in amplicon pyrosequencing studies?

Tor CARLSEN[a,], Anders Bjørnsgaard AAS[a], Daniel LINDNER[b], Trude VRÅLSTAD[a], Trond SCHUMACHER[a], Håvard KAUSERUD[a]*

→ ITS amplicons from ~100 single spore cultures (with unique tags in both ends) → sequencing

Tag switching (tag jumping, bleeding, leaking, etc.)



library sequences

Samples

	1	2	3	4	5	6	7
OTU1	0	0	0	0	0	0	0
OTU2	2	0	10000	0	0	5	0
OTU3	0	0	0	0	0	0	0
OTU4	0	0	0	0	0	0	0
OTU5	0	0	0	0	0	0	0
OTU6	0	500	0	0	0	4	0
OTU7	0	0	0	0	0	0	0
OTU8	0	0	0	0	0	0	0
OTU9	0	0	23		0	30000	0
OTU10	0	0	0	0	0	0	0

Can lead to numerous false positives!

Tag switching

MOLECULAR ECOLOGY RESOURCES
Molecular Ecology Resources (2015) doi: 10.1111/1755-0998.12402

Tag jumps illuminated – reducing sequence-to-sample misidentifications in metabarcoding studies

IDA BÆRHOLM SCHNELL,*† KRISTINE BOHMANN*‡ and M. THOMAS P. GILBERT*§
*Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, 1350 Copenhagen K, Denmark, †Center for Zoo and Wild Animal Health, Copenhagen Zoo, 2000 Frederiksberg, Denmark, ‡School of Biological Sciences, University of Bristol, Bristol BS8 1UG, UK, §Trace and Environmental DNA Laboratory, Department of Environment and Agriculture, Curtin University, Perth, Western Australia 6102, Australia

Nucleic Acids Research, 2015 1
doi: 10.1093/nar/gkv107

Accurate multiplexing and filtering for high-throughput amplicon-sequencing

Esling Philippe[1,2,*], Lejzerowicz Franck[1] and Pawlowski Jan[1]

[1]Department of Genetics and Evolution, University of Geneva, Sciences 3, 30, Quai Ernest Ansermet, CH-1211 Geneva 4, Switzerland and [2]IRCAM, UMR 9912, Université Pierre et Marie Curie, Paris, France

Received April 27, 2014; Revised January 29, 2015; Accepted January 30, 2015

"We found that an average of 2.6% and 2.1% of sequences had tag combinations, which could be explained by tag jumping..."

Up to 28.2% of the unique sequences correspond to undetectable (critical) mistags in single- or saturated double-tagging libraries.

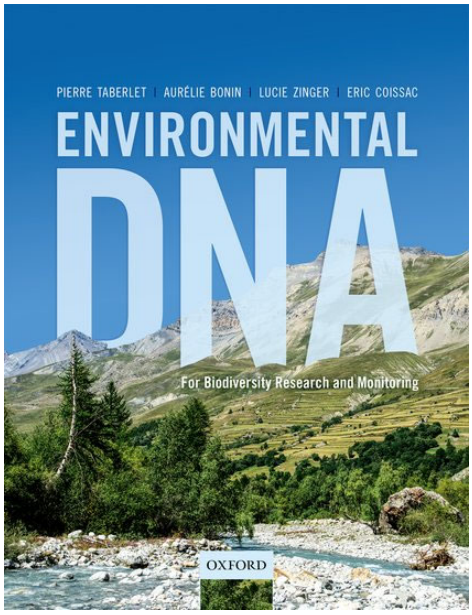
Tag switching

- The problem can be reduced or controlled for by:
  - Tagging in both ends with unique tag combinations
  - Rinse the PCR amplicons thouroughly
  - Include positive controls during PCR (mock community) → can better identify the level of switching/leakage
  - Avoid PCR step during the final library preparations steps before sequencing (i.e. when adaptors are introduced)

Tag switching

- The problem can be reduced or controlled for by:
  - Tagging in both ends with unique tag combinations
  - Rinse the PCR amplicons thouroughly
  - Avoid PCR step during the final library preparations steps before sequencing (i.e. when adaptors are introduced)
  - Include positive controls during PCR (mock community) → can better identify the level of switching/leakage

Pierre Taberlet



Tag switching

- The problem can be reduced or controlled for by:
  - Tagging in both ends with unique tag combinations
  - Rinse the PAvoid PCR step during the final library preparations steps before sequencing (i.e. when adaptors are introduced)
  - CR amplicons thouroughly
  - Include positive controls during PCR (mock community) → can better identify the level of switching/leakage

ZymoBIOMICS Microbial Community Standards

To improve the quality and reproducibility of microbiomics analyses, Zymo Research has endeavored to develop microbial reference materials. The ZymoBIOMICS Microbial Community Standard is the first commercially available standard for microbiomics and metagenomics studies. The microbial standard is a well-defined, accurately characterized mock community consisting of Gram-negative and Gram-positive bacteria and yeast with varying sizes and cell wall composition. The wide range of organisms with different properties enables characterization, optimization, and validation of lysis methods such as bead beating. It can be used as a defined input to assess the performance of entire microbiomics/metagenomic workflows, therefore enabling workflows to be optimized and validated. A mock microbial DNA community standard allows researchers to focus the optimization after the step of DNA extraction.

COMPARISON TABLE

Catalog #	Product	Size
D6300	ZymoBIOMICS Microbial Community Standard	10 Preps
D6305	ZymoBIOMICS Microbial Community DNA Standard	200 ng
D6310	ZymoBIOMICS Microbial Community Standard II (Log Distribution)	10 Preps
D6311	ZymoBIOMICS Microbial Community DNA Standard II (Log Distribution)	220ng/20µl
D6320	ZymoBIOMICS Spike-in Control I (High Microbial Load)	25 Preps
D6321	ZymoBIOMICS Spike-in Control II (Low Microbial Load)	25 Preps
D6322	ZymoBIOMICS HMW DNA Standard	5000 ng
D6323	ZymoBIOMICS Fecal Reference with TruMatrix™ Technology	10 preps
D6331	ZymoBIOMICS Gut Microbiome Standard	10 assays

Sample types

1. Biological replicates
2. Technical replicates
3. Extraction negatives
4. PCR negatives
5. Positive control (mock community)

EDITORIAL

MOLECULAR ECOLOGY WILEY

DNA metabarcoding—Need for robust experimental designs to draw sound ecological conclusions

Zinger et al. 2019. Molecular Ecology Resources

Sample types

1. Biological replicates
2. Technical replicates → Not strictly required
3. Extraction negatives
4. PCR negatives

→ Only sequence if positive amplicon?

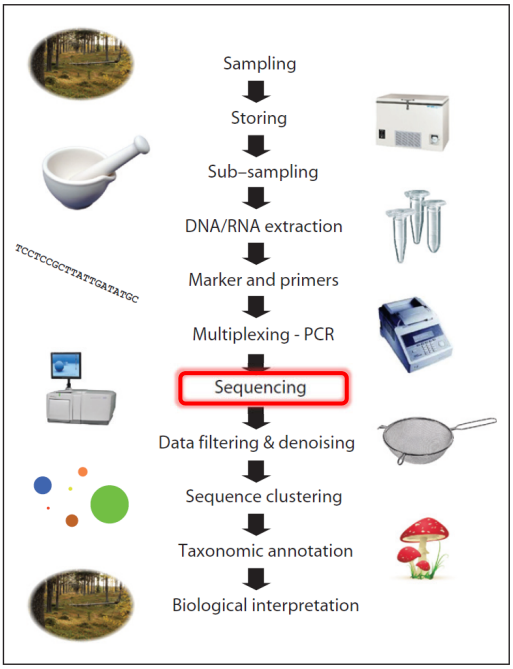
5. Positive control (mock community)

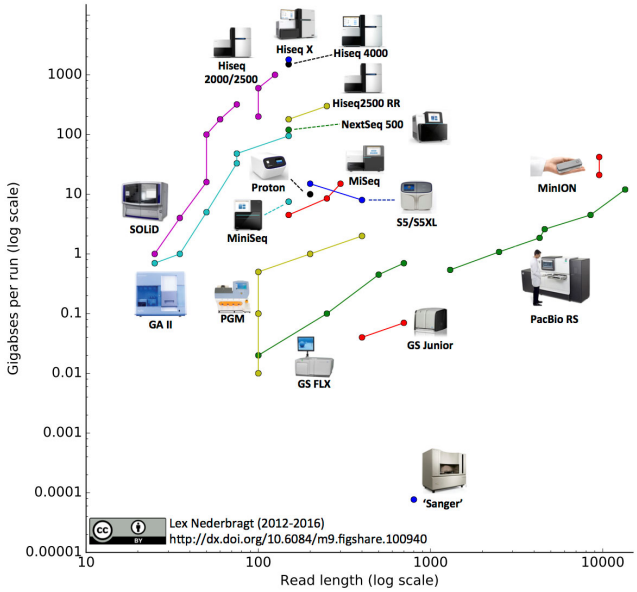
EDITORIAL

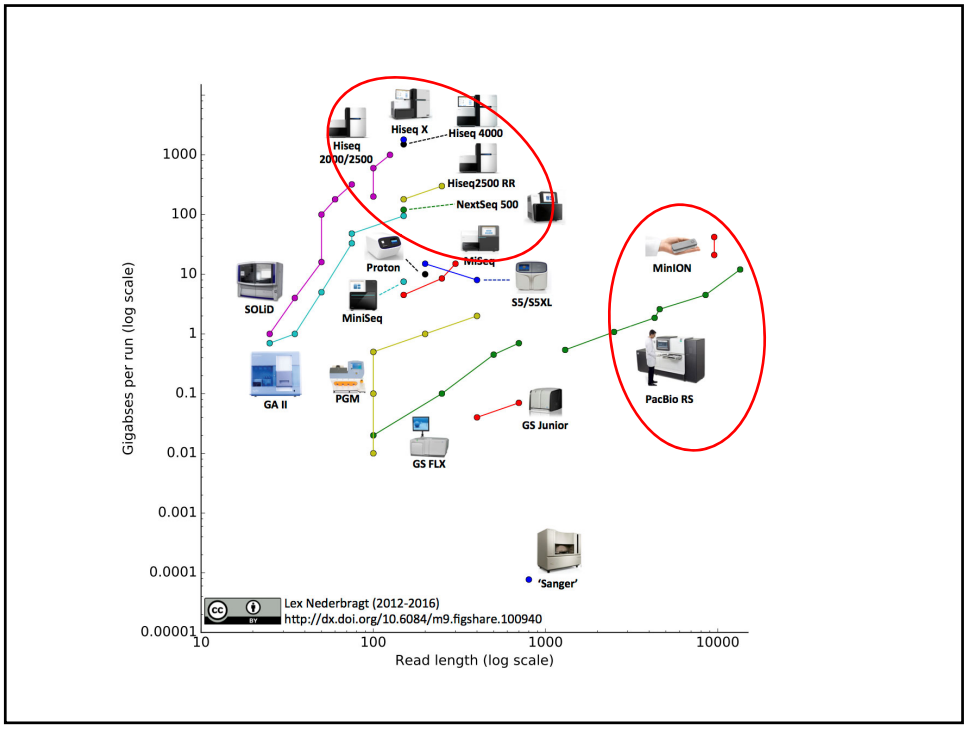
MOLECULAR ECOLOGY WILEY

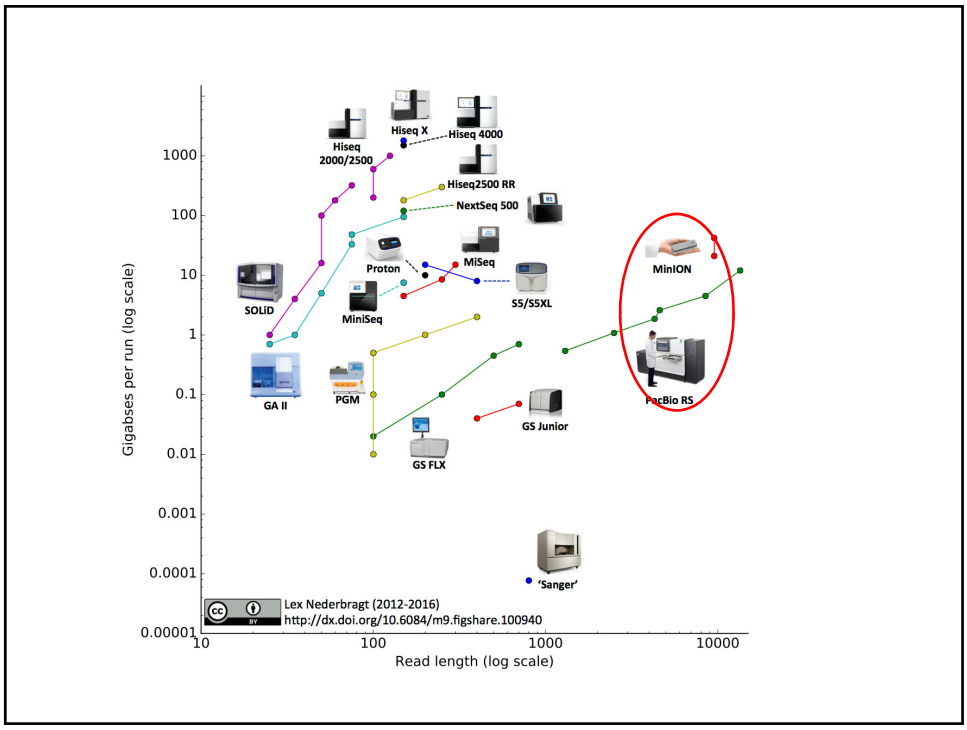
DNA metabarcoding—Need for robust experimental designs to draw sound ecological conclusions

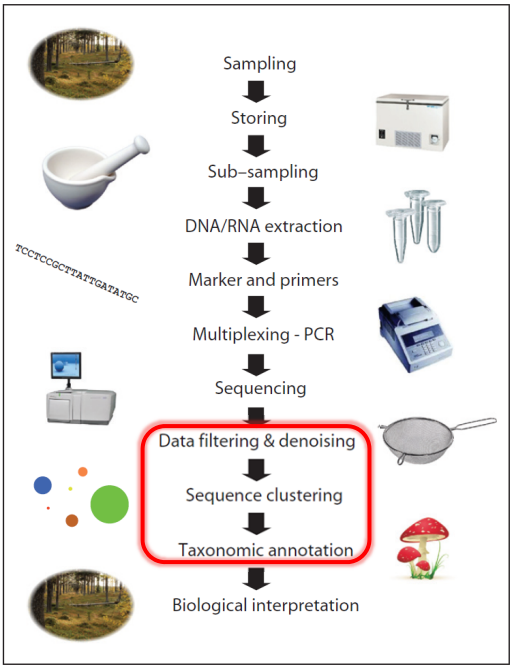
Zinger et al. 2019. Molecular Ecology Resources


Sampling
Storing
Sub–sampling
DNA/RNA extraction
TCCTCCGCTTATTGATATGC
Marker and primers
Multiplexing - PCR
Sequencing
Data filtering & denoising
Sequence clustering
Taxonomic annotation
Biological interpretation


Hiseq 2000/2500
Hiseq X
Hiseq 4000
Hiseq2500 RR
NextSeq 500
Proton
MiSeq
MiniSeq
SOLiD
S5/S5XL
MinION
GA II
PGM
GS Junior
PacBio RS
GS FLX
'Sanger'
Gigabses per run (log scale)
1000
100
10
1
0.1
0.01
0.001
0.0001
0.00001
10
100
1000
10000
Read length (log scale)
Lex Nederbragt (2012-2016)
http://dx.doi.org/10.6084/m9.figshare.100940


Hiseq 2000/2500
Hiseq X
Hiseq 4000
Hiseq2500 RR
NextSeq 500
MiSeq
Proton
S5/S5XL
MiniSeq
SOLiD
GA II
PGM
GS Junior
GS FLX
MinION
PacBio RS
'Sanger'
Gigabses per run (log scale)
Read length (log scale)
Lex Nederbragt (2012-2016)
http://dx.doi.org/10.6084/m9.figshare.100940


Hiseq 2000/2500
Hiseq X
Hiseq 4000
Hiseq2500 RR
NextSeq 500
MiSeq
Proton
S5/S5XL
MiniSeq
SOLiD
GA II
PGM
GS Junior
GS FLX
MinION
PacBio RS
'Sanger'
Gigabses per run (log scale)
Read length (log scale)
Lex Nederbragt (2012-2016)
http://dx.doi.org/10.6084/m9.figshare.100940



Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

Cleaning of tag bleeding

OTU modifications

Positive negatives

Singleton removal

Data transformation (e.g rarification)

(The order of steps depends somewhat on the pipeline/programs)

Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

Cleaning of tag bleeding

OTU modifications

Positive negatives

Singleton removal

Data transformation (e.g rarification)

Bioinformatics steps

Post processing steps

Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

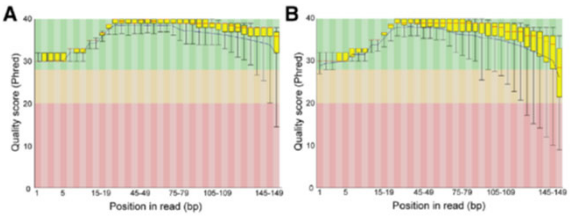
Cleaning of tag bleeding

OTU modifications

Positive negatives

Singleton removal

Data transformation (e.g rarification)



Remove

- Poor sequence quality
- Long/short sequences

Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

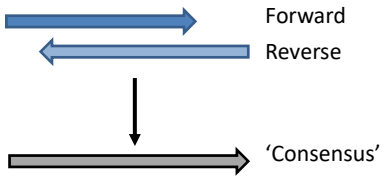
Cleaning of tag bleeding

OTU modifications

Positive negatives

Singleton removal

Data transformation (e.g rarification)



Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

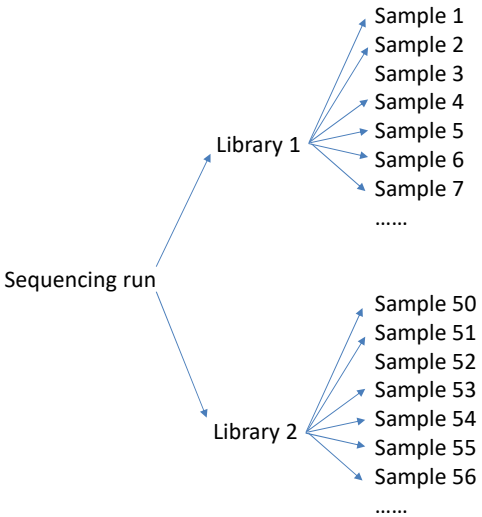
Cleaning of tag bleeding

OTU modifications

Positive negatives

Singleton removal

Data transformation (e.g rarification)



Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

Cleaning of tag bleeding

OTU modifications

Positive negatives

Singleton removal

Data transformation (e.g rarification)

(k-mer based analyses)

Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

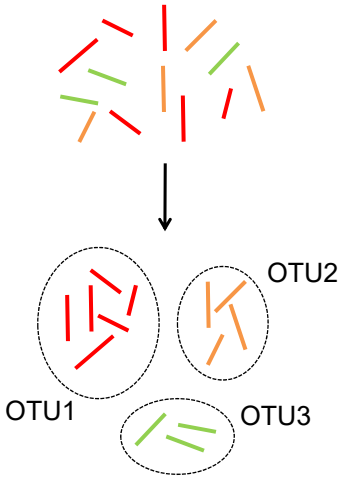
Cleaning of tag bleeding

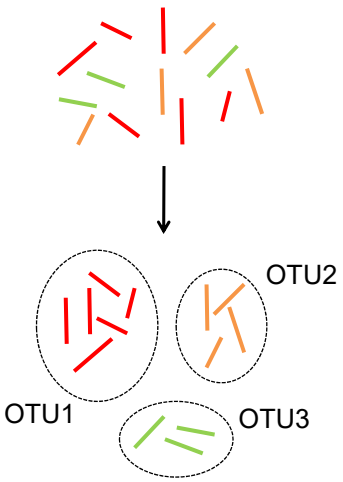
OTU modifications

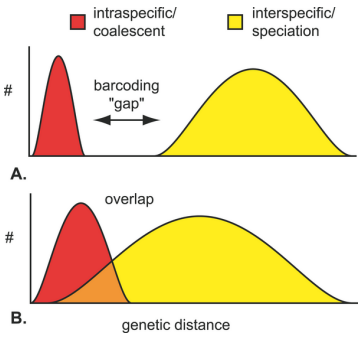
Positive negatives

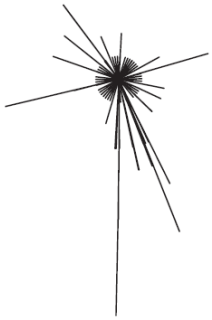
Singleton removal

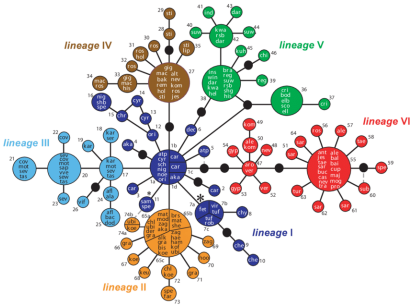
Data transformation (e.g rarification)











PCR induced errors versus natural sequence variation?

Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

Cleaning of tag bleeding

OTU modifications

Positive negatives

Singleton removal

Data transformation (e.g rarification)

de novo versus closed (reference based) OTU construction?

Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

Cleaning of tag bleeding

OTU modifications

Positive negatives

Singleton removal

Data transformation (e.g rarification)

de novo v~~ersus closed (reference based)~~ OTU construction?

Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

Cleaning of tag bleeding

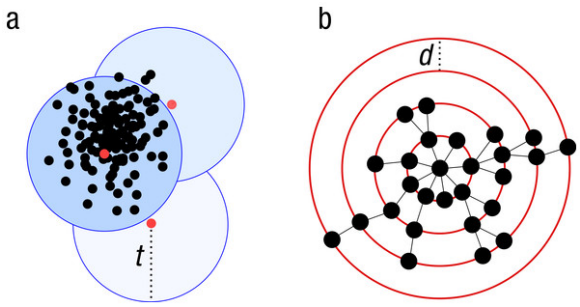
OTU modifications

Positive negatives

Singleton removal

Data transformation (e.g rarification)

Many different clustering approaches



Mahé et al. 2014

Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

Cleaning of tag bleeding

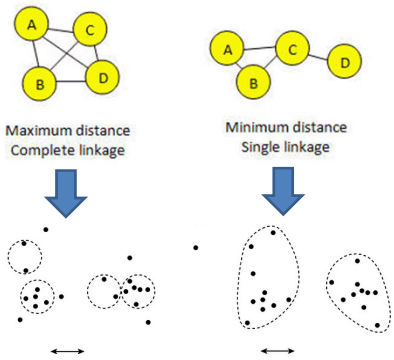
OTU modifications

Positive negatives

Singleton removal

Data transformation

Many different clustering approaches



Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

Cleaning of tag bleeding

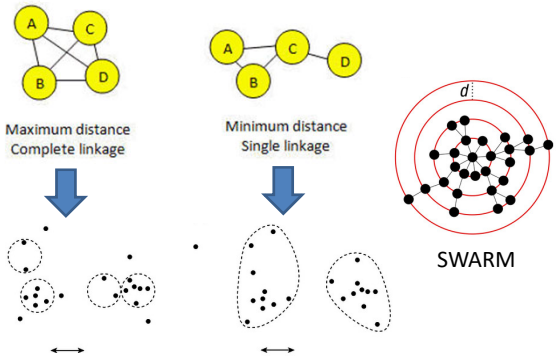
OTU modifications

Positive negatives

Singleton removal

Data transformation

Many different clustering approaches



Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

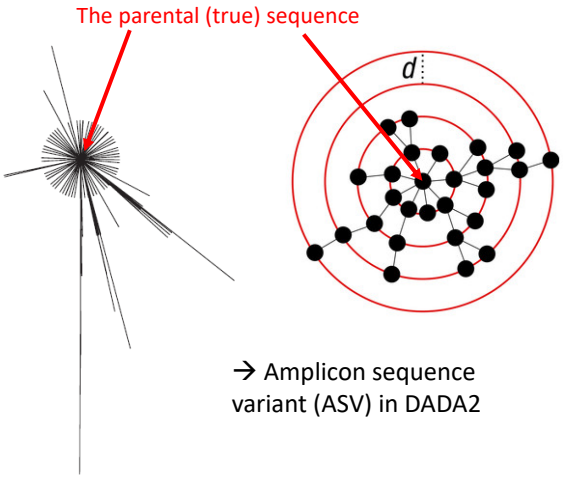
Cleaning of tag bleeding

OTU modifications

Positive negatives

Singleton removal

Data transformation (e.g rarification)



→ Amplicon sequence variant (ASV) in DADA2

Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

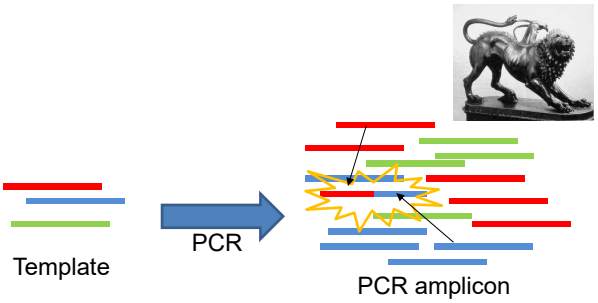
Cleaning of tag bleeding

OTU modifications

Positive negatives

Singleton removal

Data transformation (e.g rarification)



De novo versus reference based chimera checking

Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

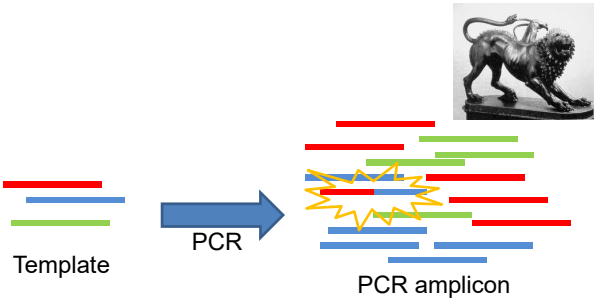
Cleaning of tag bleeding

OTU modifications

Positive negatives

Singleton removal

Data transformation (e.g rarification)



De novo ~~versus reference based~~ chimera checking

The level of chimeric sequences depends on how variable the marker is! → Be aware of false positives in the tests

Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

Cleaning of tag bleeding

OTU modifications

Positive negatives

Singleton removal

Data transformation (e.g rarification)

Simple matching (blast) → probabilistic assignment (protax)

Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

Cleaning of tag bleeding

OTU modifications

Positive negatives

Singleton removal

Data transformation (e.g rarification)

FROM THE COVER

MOLECULAR ECOLOGY RESOURCES WILEY

Assessment of current taxonomic assignment strategies for metabarcoding eukaryotes

Jose S. Hleap[1,2,3] | Joanne E. Littlefair[1,4] | Dirk Steinke[5] | Paul D. N. Hebert[5] | Melania E. Cristescu[1]

Hleap et al. 2021. Mol Ecol Resources

Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

Cleaning of tag bleeding

OTU modifications

Positive negatives

Singleton removal

Data transformation (e.g rarification)

Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

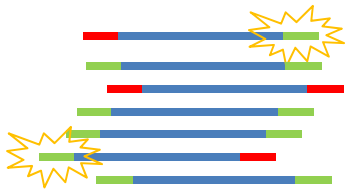
Cleaning of tag bleeding

OTU modifications

Positive negatives

Singleton removal

Data transformation

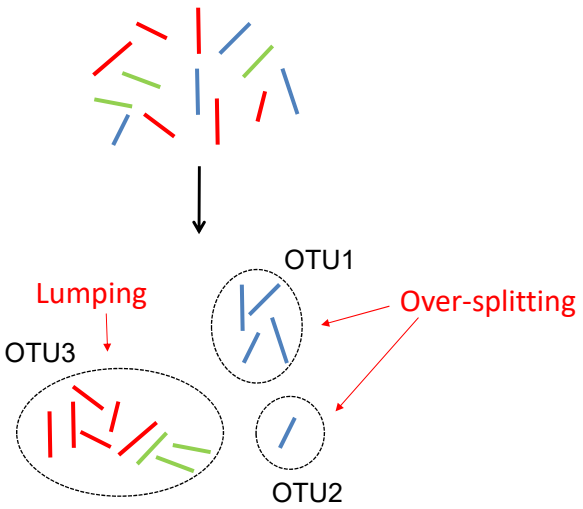


Samples

	1	2	3	4	5	6	7
OTU1	0	0	0	0	0	0	0
OTU2	2	0	10000	0	0	5	0
OTU3	0	0	0	0	0	0	0
OTU4	0	0	0	0	0	0	0
OTU5	0	0	0	0	0	0	0
OTU6	0	500	0	0	0	4	0
OTU7	0	0	0	0	0	0	0
OTU8	0	0	0	0	0	0	0
OTU9	0	0	23		0	30000	0
OTU10	0	0	0	0	0	0	0

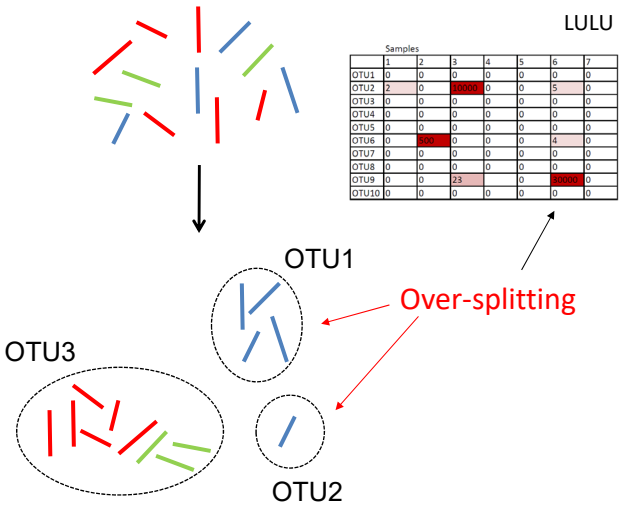
Bioinformatics – main steps

Quality control
Produce contigs
Demultiplexing
Dereplication
OTU construction
Chimera checking
Taxonomic annotation
Removal of non-target organisms
Cleaning of tag bleeding
OTU modifications
Positive negatives
Singleton removal
Data transformation



Bioinformatics – main steps

Quality control
Produce contigs
Demultiplexing
Dereplication
OTU construction
Chimera checking
Taxonomic annotation
Removal of non-target organisms
Cleaning of tag bleeding
OTU modifications
Positive negatives
Singleton removal
Data transformation



Bioinformatics – main steps

Quality control

Produce contigs

ARTICLE

DOI: 10.1038/s41467-017-01312-x OPEN

Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates

Tobias Guldberg Frøslev[1,2], Rasmus Kjøller[1], Hans Henrik Bruun[1], Rasmus Ejrnæs[3], Ane Kirstine Brunbjerg[3], Carlotta Pietroni[2] & Anders Johannes Hansen[2]

LULU

	Samples						
	1	2	3	4	5	6	7
OTU1	0	0	0	0	0	0	0
OTU2	2	0	10000	0	0	5	0
OTU3	0	0	0	0	0	0	0
OTU4	0	0	0	0	0	0	0
OTU5	0	0	0	0	0	0	0
OTU6	0	900	0	0	0	4	0
OTU7	0	0	0	0	0	0	0
OTU8	0	0	0	0	0	0	0
OTU9	0	0	23		0	300000	0
OTU10	0	0	0	0	0	0	0

Taxonomic annotation

Removal of non-target organisms

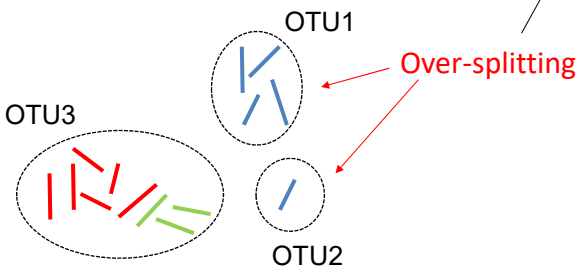
Cleaning of tag bleeding

OTU modifications

Positive negatives

Singleton removal

Data transformation



Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

Cleaning of tag bleeding

OTU modifications

Positive negatives

Singleton removal

Data transformation

External or internal contaminants?

RESEARCH ARTICLE
Novel Systems Biology Techniques

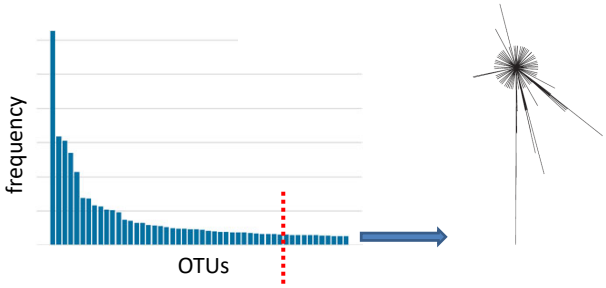
Quantifying and Understanding Well-to-Well Contamination in Microbiome Research

Jeremiah J. Minich,[a] Jon G. Sanders,[b] Amnon Amir,[b] Greg Humphrey,[b] Jack A. Gilbert,[c,d] Rob Knight[b,e,f,g]

Should be very careful how you treat positive negatives!!

Bioinformatics – main steps

Quality control
Produce contigs
Demultiplexing
Dereplication
OTU construction
Chimera checking
Taxonomic annotation
Removal of non-target organisms
Cleaning of tag bleeding
OTU modifications
Positive negatives
Singleton removal
Data transformation



What is a 'singleton'? → Depends on your sequencing depth and quality of your data. Should also take study aim into consideration

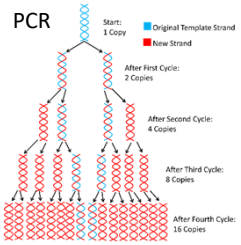
Bioinformatics – main steps

Quality control
Produce contigs
Demultiplexing
Dereplication
OTU construction
Chimera checking
Taxonomic annotation
Removal of non-target organisms
Cleaning of tag bleeding
OTU modifications
Positive negatives
Singleton removal
Data transformation

Samples

	1	2	3	4	5	6	7
OTU1	0	0	0	0	0	0	0
OTU2	2	0	10000	0	0	5	0
OTU3	0	0	0	0	0	0	0
OTU4	0	0	0	0	0	0	0
OTU5	0	0	0	0	0	0	0
OTU6	0	500	0	0	0	4	0
OTU7	0	0	0	0	0	0	0
OTU8	0	0	0	0	0	0	0
OTU9	0	0	23		0	30000	0
OTU10	0	0	0	0	0	0	0

PCR



Be careful with resampling and transformations!

Check the effect from various data treatments options on the results!

Depends on the study aims!

The importance of controls

1. Biological replicates
2. Technical replicates
3. Extraction negatives
4. PCR negatives
5. Positive control (mock community)

→ Different purposes

Conclusions

- Which methods to use? → No general answer – it is context dependent. You must argue for your choices!