Welcome to the course *Bioinformatics for Environmental Sequencing (BIO9905MERG1)*

Learning goals

- Have a conceptual understanding of the various steps during bioinformatics analyses of metabarcoding data .
- Understand and argue for why you make your choices and use various bioinformatics approaches in different situations. Learn to be critical towards DNA metabarcoding data and how to evaluate them.
- Develop basic knowlege in important programs, like cutadapt, DADA2, SWARM and VSEARCH.
- Obtain knowledge about long-read metabarcoding and phylogenetic placement.
- Obtain insight in possible downstream analyses.
- By discussing with your fellow students – learn about different DNA metabarcoding projects.

Organization

- Mix of lectures + hands on sessions + discussions
- Basic → advanced. High variation in skills/experience. Experienced students should preferably assist less experienced students
- We use a slack as channel for micro-communication
- Course information at https://github.com/krabberod/BIO9905MERG1_V25
- Hands-on sessions in R and Google colab
- Report: Instructions provided at github
- A rather complex course to organize - please have patience..
- Be active and engaged!

Program and teachers

Day	Time (start)	Topic	Responsible
Monday	09:00-11:00	Introduction to DNA metabarcoding + hello	Håvard Kauserud
	11:00-12:00	Introduction to DNA sequencing techniques and data formats	Anders Krabberød
	12:00-13:00	Lunch break	
	13:00-14:00	Get to know each other (group work)	Håvard Kauserud
	14:00-16:30	Introduction to Linux, Google Colab, R, cutadapt, etc.	Ramiro Logares/Anders Krabberød
	17:00-	Pizza for all participants	
Tuesday	09:00-12:00	DADA2	Anders Krabberød/Ramiro Logares
	12:00-13:00	Lunch break	
	13.00-14.00	Case study: Long-read metabarcoding of tropical marine protist co	Denise Rui Ying Ong
	14:00-17:00	DADA2	Anders Krabberød/Ramiro Logares
Wednesday	09:00-10:00	Introduction to long-read DNA metabarcoding	Embla Stokke
	10:00-11:00	Introduction to VSEARCH (and SWARM)	Torbjørn Rognes
	11:00-12:00	The VSEARCH pipeline, hands on session	Ramiro Logares/Anders Krabberød
	12:00-13:00	Lunch break	
	13:00-14:00	The VSEARCH pipeline, hands on session, continuation	Ramiro Logares/Anders Krabberød
	14:00-15:00	Introduction to LULU/MUMU	Frédéric Mahé (zoom)
	15:00-16:00	Data cleanup (contamination, etc.) + discussions	Håvard Kauserud
Thursday	09:00-10:00	OTUs, ASVs and phylospecies	Micah Dunthorn
	10:00-12:00	PR2, metaPR2 + other databases	Daniel Vaulot/Anders Krabberød
	12:00-13:00	Lunch break	
	13:00-15:00	Phylogenetic placement/binning of HTS data	Lucas Czech (zoom)
	15:00-16:00	Case study long read metabarcoding	Ella Thoen
Friday	09:00-10:00	Taxonomic annotation	Marie Davey (zoom)
	10:00-11:00	Metacoder	Ella Thoen
	11:00-12:00	DNA metabarcoding and contamination	Kristine Bohmann (zoom)
	12:00-13:00	Lunch break	
	13:00-14:00	Downstream analyses: Multivariate analyses	Ramiro Logares
	14:00-15:00	Downstream analyses: Network inferences	Anders Krabberød
	15:00-16:00	Q&A session	Multiple teachers

Bioinformatics workflow

- QC Sequencing Results
- Demultiplexing
- Quality control, filtering, trimming
- Dereplication
- Denoising / OTU clustering
- Chimera removal
- OTU table construction
- Taxonomic assignment
- Removal of non-targets
- Normalization or rarefaction
- Downstream analysis and plotting

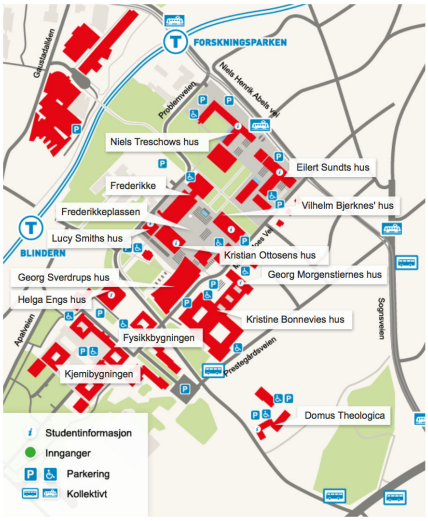
Program and teachers – who are we?

Day	Time (start)	Topic	Responsible
Monday	09:00-11:00	Introduction to DNA metabarcoding + hello	Håvard Kauserud
	11:00-12:00	Introduction to DNA sequencing techniques and data formats	Anders Krabberød
	12:00-13:00	Lunch break	
	13:00-14:00	Get to know each other (group work)	Håvard Kauserud
	14:00-16:30	Introduction to Linux, Google Colab, R, cutadapt, etc.	Ramiro Logares/Anders Krabberød
	17:00-	Pizza for all participants	
Tuesday	09:00-12:00	DADA2	Anders Krabberød/Ramiro Logares
	12:00-13:00	Lunch break	
	13.00-14.00	Case study: Long-read metabarcoding of tropical marine protist co	Denise Rui Ying Ong
	14:00-17:00	DADA2	Anders Krabberød/Ramiro Logares
Wednesday	09:00-10:00	Introduction to long-read DNA metabarcoding	Embla Stokke
	10:00-11:00	Introduction to VSEARCH (and SWARM)	Torbjørn Rognes
	11:00-12:00	The VSEARCH pipeline, hands on session	Ramiro Logares/Anders Krabberød
	12:00-13:00	Lunch break	
	13:00-14:00	The VSEARCH pipeline, hands on session, continuation	Ramiro Logares/Anders Krabberød
	14:00-15:00	Introduction to LULU/MUMU	Frédéric Mahé (zoom)
	15:00-16:00	Data cleanup (contamination, etc.) + discussions	Håvard Kauserud
Thursday	09:00-10:00	OTUs, ASVs and phylospecies	Micah Dunthorn
	10:00-12:00	PR2, metaPR2 + other databases	Daniel Vaulot/Anders Krabberød
	12:00-13:00	Lunch break	
	13:00-15:00	Phylogenetic placement/binning of HTS data	Lucas Czech (zoom)
	15:00-16:00	Case study long read metabarcoding	Ella Thoen
Friday	09:00-10:00	Taxonomic annotation	Marie Davey (zoom)
	10:00-11:00	Metacoder	Ella Thoen
	11:00-12:00	DNA metabarcoding and contamination	Kristine Bohmann (zoom)
	12:00-13:00	Lunch break	
	13:00-14:00	Downstream analyses: Multivariate analyses	Ramiro Logares
	14:00-15:00	Downstream analyses: Network inferences	Anders Krabberød
	15:00-16:00	Q&A session	Multiple teachers

Who are you?

Practical information

- How to access to the third floor? IBV people, please help out!
- Toilets
- Coffee and tea outside, as well as in the fourth floor
- Food/lunch opportunities
- Pizza on Monday, after the course?
- Dinner and beer/drinks at Oslo Streetfood, Wednesday 6pm, please register

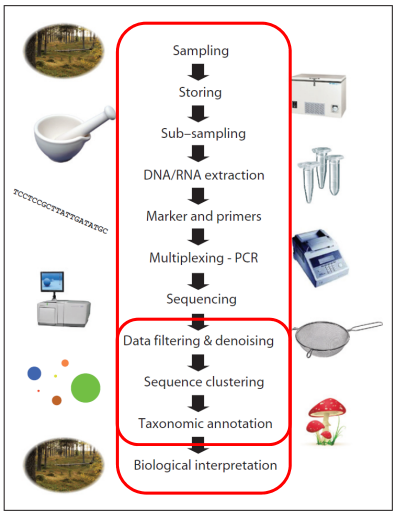


Welcome to the course *Bioinformatics for Environmental Sequencing (BIO9905MERG1)*

ForBio
Forskerskole i Biosystematikk
Research School in Biosystematics

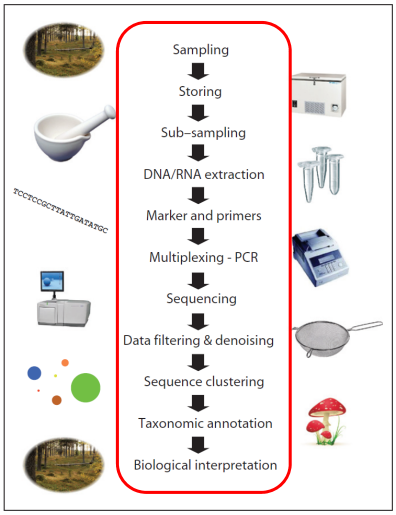
Introduction to metabarcoding

- Introduce key steps, including sampling and DNA wet lab work
- Emphasise important choices to be made in light of your marker and study organisms
- Explain important terms
- Introduce some literature
- More in-depth information in later talks



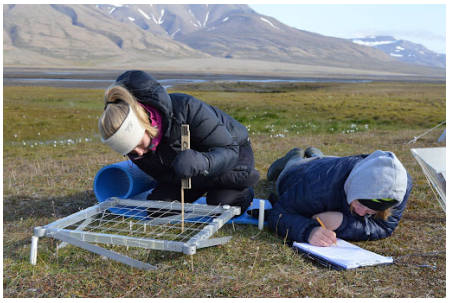
Introduction to metabarcoding

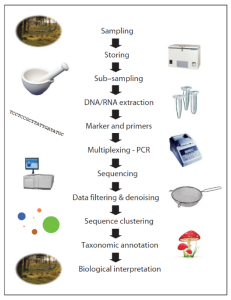
- Introduce key steps, including sampling and DNA wet lab work
- Emphasise important choices to be made in light of your marker and study organisms
- Explain important terms
- Introduce some literature
- More in-depth information in later talks



Introduction to metabarcoding

- *«Metabarcoding is very simple, and very complex»*

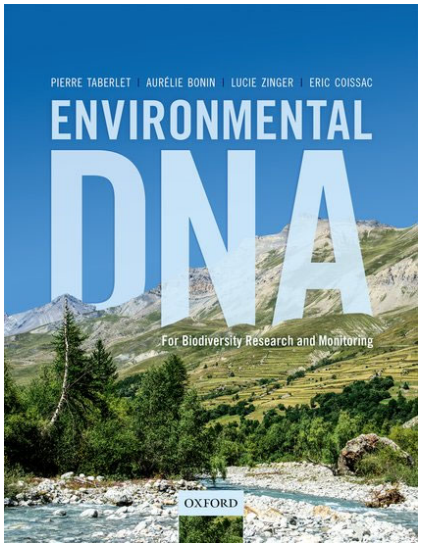




Pierre Taberlet

Introduction to metabarcoding

Pierre Taberlet



Some important but confusing terms

ASVs

amplicon sequencing

OTUs

DNA metabarcoding

metatranscriptomics

eDNA

rDNA phylotyping

metagenetics

metagenomics

microbiome analysis

environmental sequencing

MOTUs

community profiling

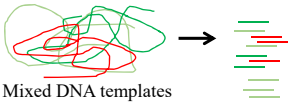
marker gene analysis

Some important but confusing terms

- DNA barcoding

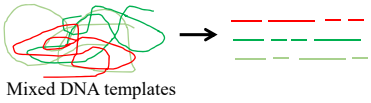
Sequence variation in a single locus (e.g. ITS) in a single specimen

- Metabarcoding



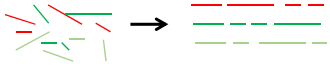
Sequence variation in a single locus (e.g. ITS) in a community

- Metagenomics



Genome wide sequence variation in a community

- Metatranscriptomics



cDNA sequence variation in a community

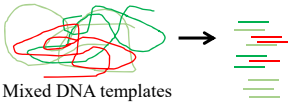
Some important but confusing terms

- DNA barcoding

Sequence variation in a single locus (e.g. ITS) in a single specimen

Species identification

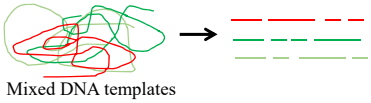
- Metabarcoding



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

Who are there?

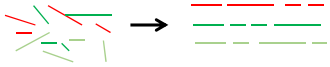
- Metagenomics



Genome wide sequence variation in a community

Which genes (and who) are there?

- Metatranscriptomics



cDNA sequence variation in a community

Who are active and doing what?

Some important but confusing terms

ASVs

amplicon sequencing

OTUs

DNA metabarcoding

metatranscriptomics

eDNA

rDNA phylotyping

metagenomics

metagenetics

microbiome analysis

MOTUs

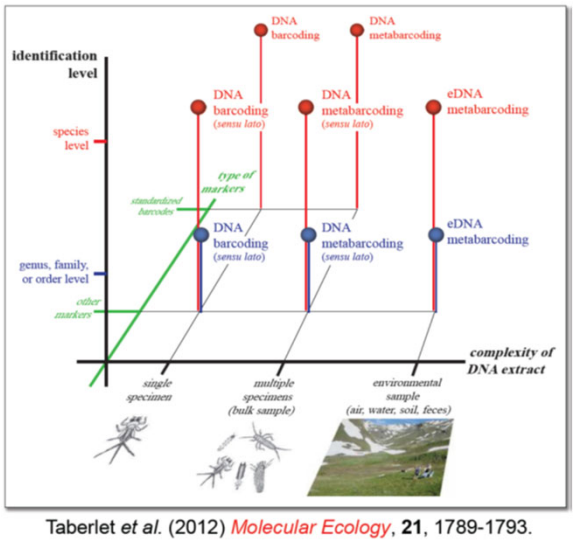
environmental sequencing

community profiling

marker gene analysis

Terminology

Pierre Taberlet



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

Some important but confusing terms

In most cases, there are some uncertainty associated with the taxonomic identification

ASVs

amplicon sequencing

OTUs

DNA metabarcoding

metatranscriptomics

eDNA

rDNA phylotyping

metagenetics

metagenomics

MOTUs

environmental sequencing

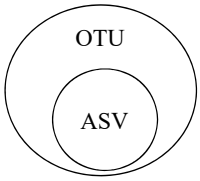
microbiome analysis

marker gene analysis

community profiling

Some important but confusing terms

«An ASV is always and OTU, but an OTU is not always an ASV»
Anders K. Krabberød



ASVs

amplicon sequencing

OTUs

DNA metabarcoding

metatranscriptomics

eDNA

rDNA phylotyping

metagenetics

metagenomics

MOTUs

environmental sequencing

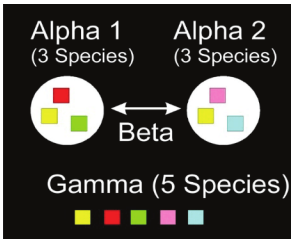
microbiome analysis

marker gene analysis

community profiling

DNA metabarcoding

- Research questions
  - Who are there?
  - How many operational taxonomic units are present (alpha/gamma diversity)?
  - How do communities differ in composition (beta diversity)?
  - Which processes and drivers are shaping the communities?
  - Co-occurrence patterns → Interactions
- Provides qualitative, proportional data, not absolute quantitative data
  - Can say who are common and rare within the samples (relative abundances), but not in more absolute terms how common or rare. «Semi-quantitative».



From «wild west» towards an established approach

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

EDITORIAL

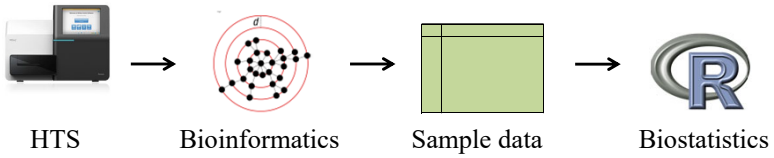
MOLECULAR ECOLOGY WILEY

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

Zinger et al. 2019, Molecular Ecology Resources

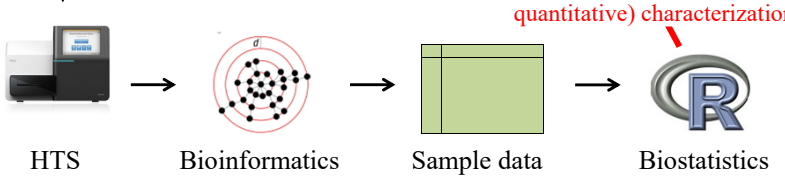
Secondary phase
- Established scientific approach with a set of widely accepted guidelines

General workflow

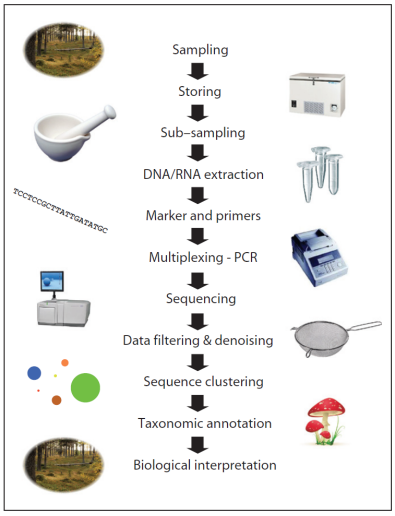


General workflow

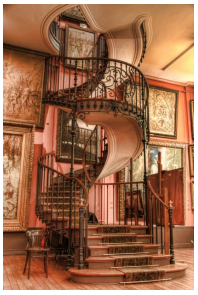
Qualitative (and semi-quantitative) characterization



DNA metabarcoding - many steps



Lindahl et al. 2013



Many steps…

… to go wrong

Many good review papers

EDITORIAL

MOLECULAR ECOLOGY WILEY

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

Zinger et al. 2019. Molecular Ecology Resources

Review

Cell

Environmental DNA for wildlife biol… and biodiversity monitoring

Kristine Bohmann[1,2*], Alice Evans[3*], M. Thomas P. Gilbert[1,4], Gary R. Carvalho[3], Simon Creer[3], Michael Knapp[3], Douglas W. Yu[5,6], and Mark de Bruyn[3]

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]

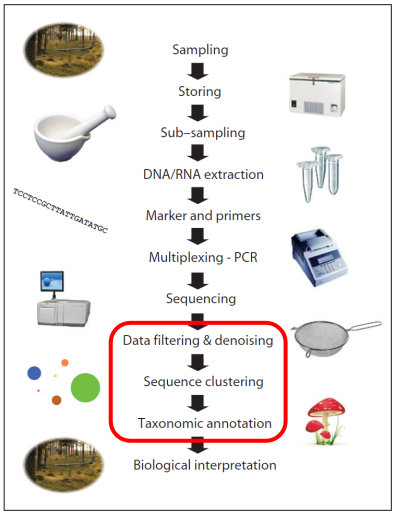
NEWS AND VIEWS

OPINION

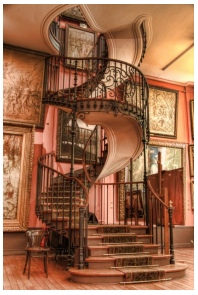
Towards exhaustive community ecology via DNA metabarcoding

Gentile Francesco Ficetola[1,2] | Pierre Taberlet[2,3]

DNA metabarcoding - many steps



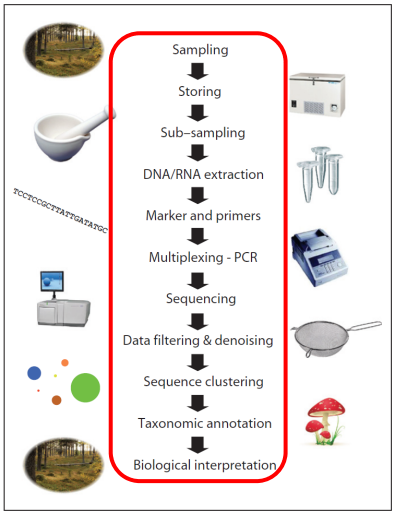
Lindahl et al. 2013



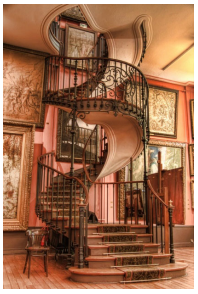
Many steps…

… to go wrong

DNA metabarcoding - many steps



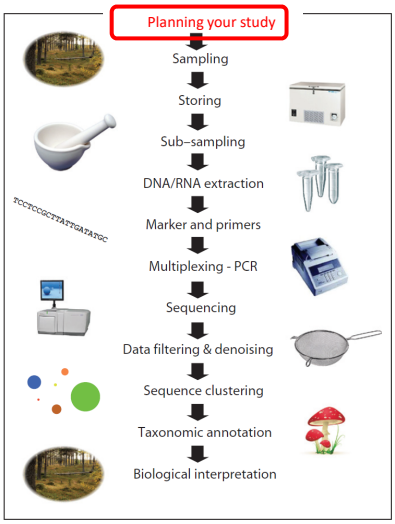
Lindahl et al. 2013



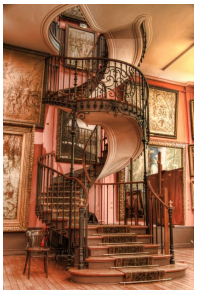
Many steps…

… to go wrong

DNA metabarcoding - many steps



Lindahl et al. 2013



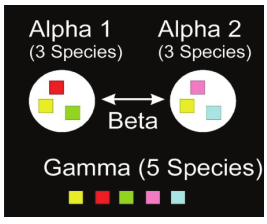
Many steps…

… to go wrong

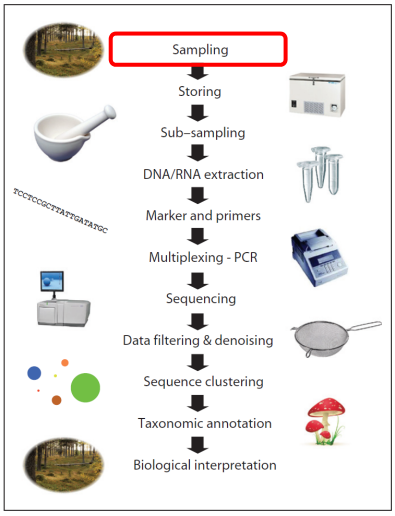
If new study system – conduct a pilot?

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

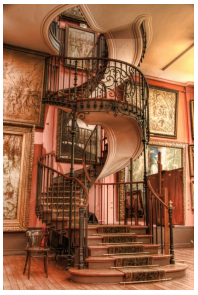
→ Depends on the alpha, beta and gamma diversity, that you might not know anything about..



DNA metabarcoding - many steps



Lindahl et al. 2013



Many steps…

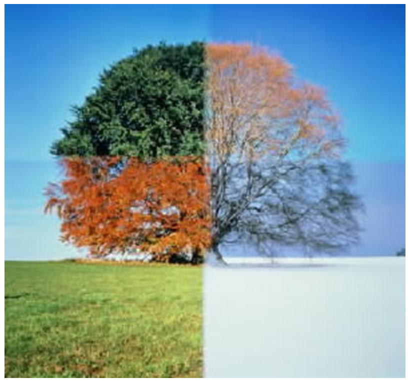
… to go wrong

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?

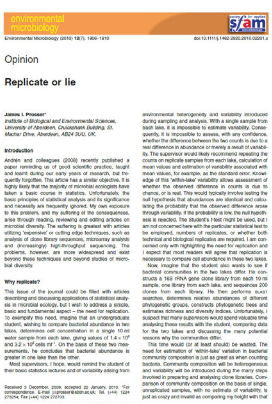
Fig. 5. What is the optimal relationship between primary sample size and the analytical sample volume (insert) and how can it come about? When sample size increases one can intuitively understand that the sample becomes more representative. But at the same time, today's analytical volumes continue to 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

«Replicate or lie»

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[7] | 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]

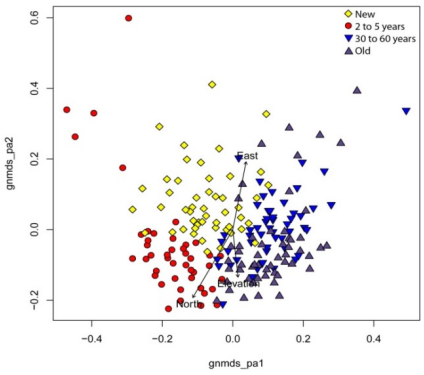
Abstract

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



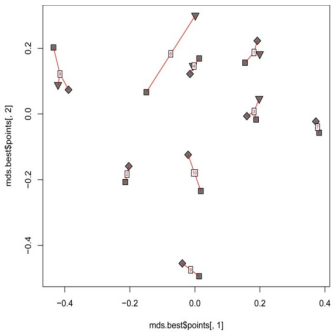
Davey *et al.* 2014. FEMS Microbial Ecology

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?

NB! Communities with low DNA content

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

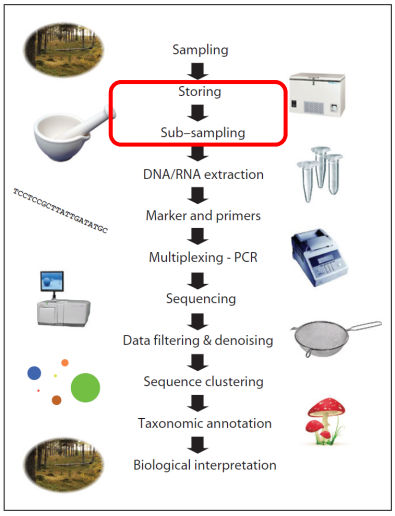


Davey *et al.* 2014. FEMS Microbial Ecology

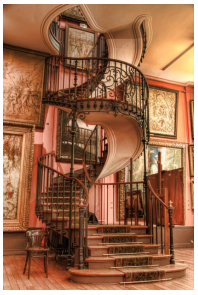
Different sample types

1. Biological replicates
2. Technical replicates

DNA metabarcoding - many steps



Lindahl et al. 2013



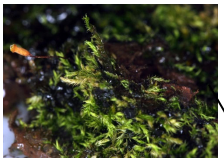
Many steps…

… to go wrong

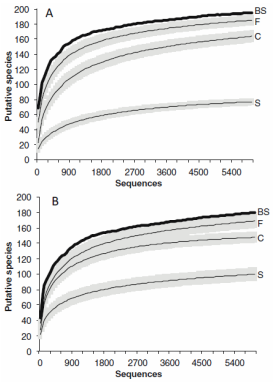
Storage

- 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

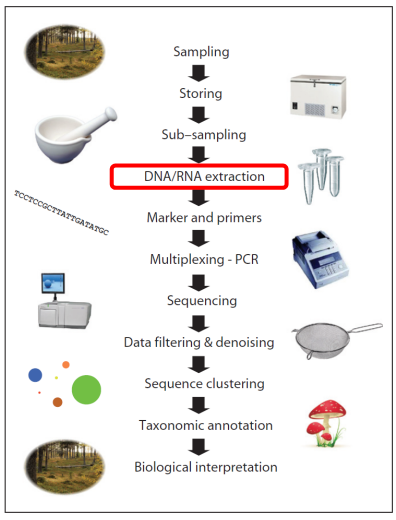
Sub-sampling and homogenization

- Protocols for nucleic acid extraction are normally based on small amounts
- Field samples should preferable be 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.

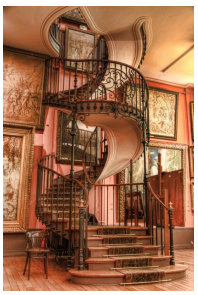
Homogenization of samples

- 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!

DNA metabarcoding - many steps



Lindahl et al. 2013



Many steps…

… to go wrong

DNA extraction

DNA extraction

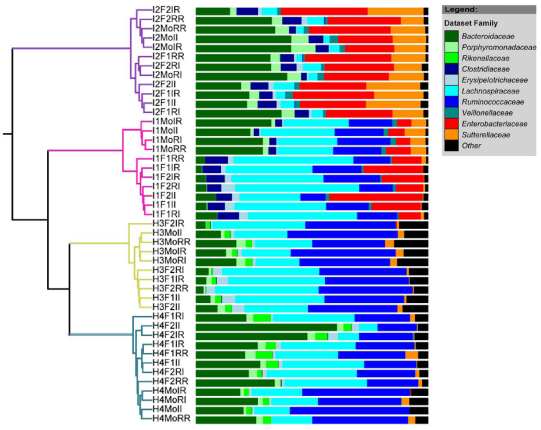
- Should yield high and uniform amounts of DNA
- Concentration of PCR inhibitors (in e.g. soil) should be minimized
- Same protocol for all samples!
- If no proper literature are available on your study system → conduct a pilot?!

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

DNA extraction

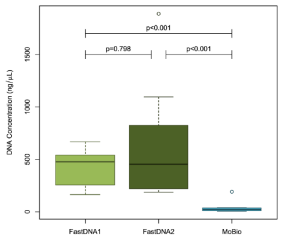
The Impact of Different DNA Extraction Kits and Laboratories upon the Assessment of Human Gut Microbiota Composition by 16S rRNA Gene Sequencing

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*†]



Key to sample names

Patient or control	Extraction method	Site of DNA extraction	Site of PCR
I1 I2 (IBD patients) H3 H4 (healthy controls) Clades within the tree are coloured acccording to donor	F1, F2 (FastDNA methods 1 and 2) Mo (MoBio PowerSoil)	I (Institute of Medical Sciences) R (Rowett Institute)	



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

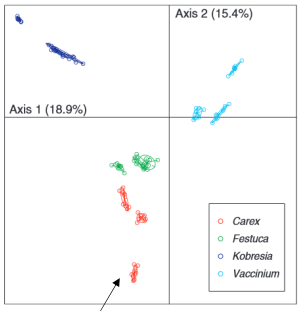
DNA extraction

- Both extracellular and intracelluar DNA are often co-extracted from many substrates, e.g. soil and water samples
- Method for extraction (mainly) 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

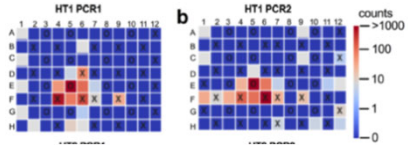


Contamination during extration

mSystems
RESEARCH ARTICLE
Novel Systems Biology Techniques

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

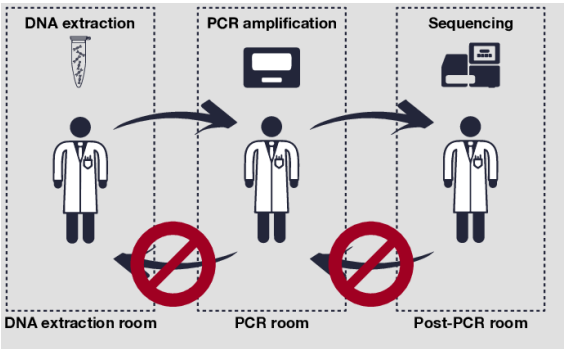
Jeremiah J. Minich,[a] Jon G. Sanders,[b] Amnon Amir,[b] Greg Humphrey,[b] Jack A. Gilbert,[c,d] Rob Knight[b,e,f]



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

DNA extraction

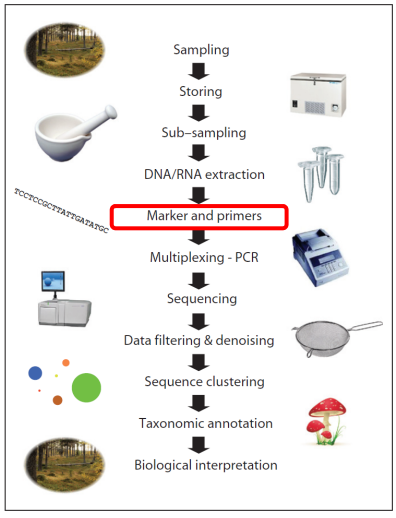


- Extraction biases → Extraction negatives!

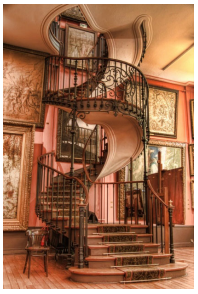
Sample types

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

DNA metabarcoding - many steps



Lindahl et al. 2013

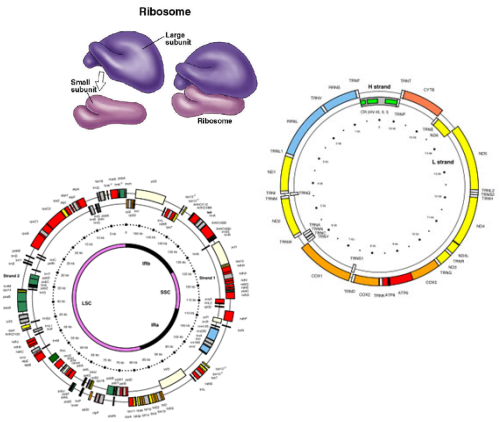


Many steps…

… to go wrong

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



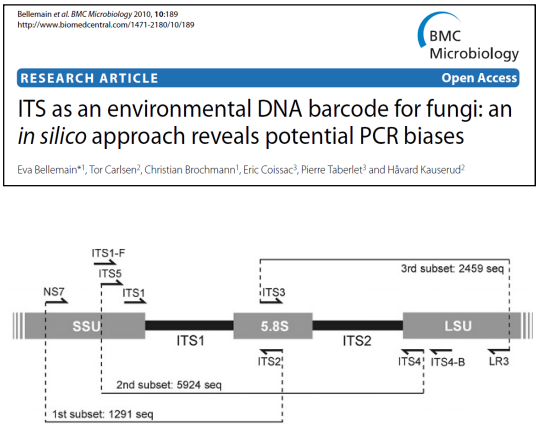
Markers in DNA metabarcoding

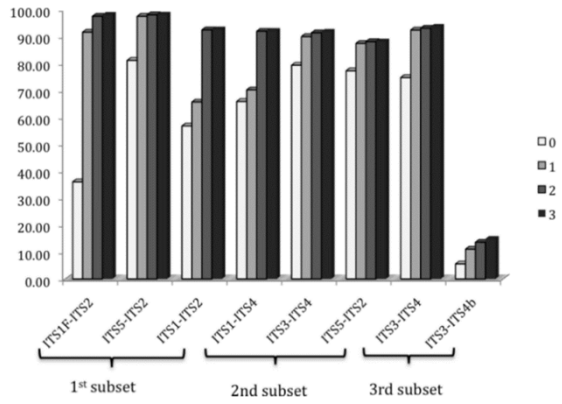
- The ideal marker should:
  - Have primer sites that are shared by all target organisms
  - Be easy to amplify
  - Be of appropriate length for efficient amplification and sequencing
  - Be of similar length
  - No intragenomic variation (i.e. no paralogs)
  - Similar number of copies
  - Be possible 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. 2010, BMC Microbiology

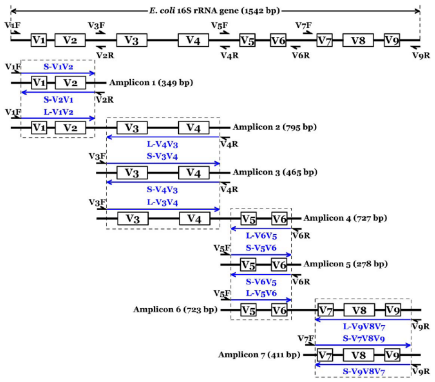
16S and primer bias

OPEN ACCESS Freely available online PLOS ONE

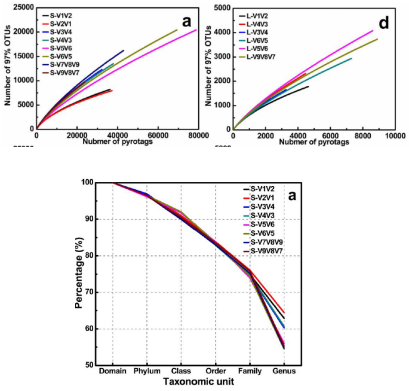
Biased Diversity Metrics Revealed by Bacterial 16S Pyrotags Derived from Different Primer Sets

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

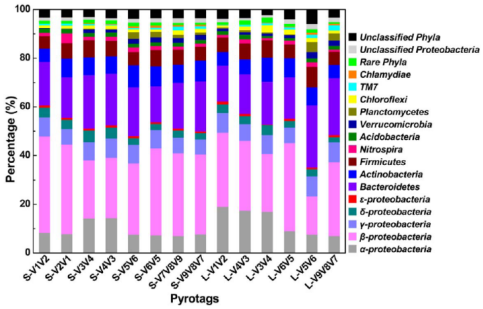
OPEN ACCESS Freely available online PLOS ONE

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

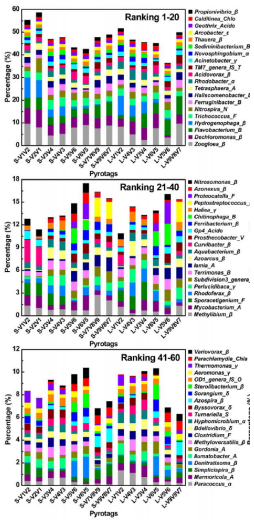
Lin Cai[1], Lin Ye[1], Amy Hin Yan Tong[2], Si Lok[2], Tong Zhang[1]*

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

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

The PR2 primer database!

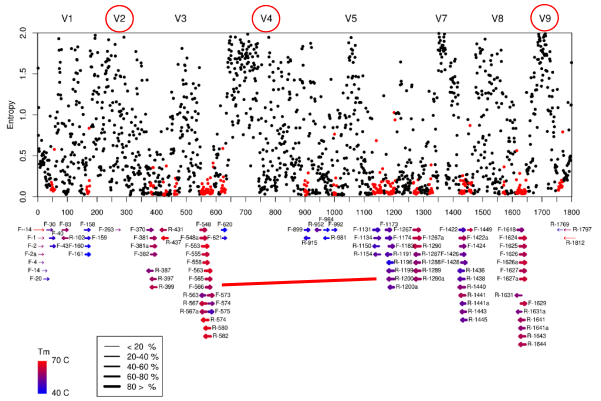


Figure 1. Nucleotide variability in 18S rRNA genes. Shannon entropy values of all eukaryotic alignment positions from the SILVA database along the 18S rRNA gene of *Saccharomyces cerevisiae*. Red dots mark consecutive nucleotide positions where at least 90% of ≥10 nt have entropy values lower than 0.2. 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

in silico primer evaluation

Good programs to design and evalute primers: EcoPrimers and EcoPCR

Published online 19 September 2011 *Nucleic Acids Research, 2011, Vol. 39, No. 21* **e145**
doi:10.1093/nar/gkr732

ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis

Tiayyba Riaz[1], Wasim Shehzad[1], Alain Viari[2], François Pompanon[1], Pierre Taberlet[1] and Eric Coissac[1,*]

[1]Laboratoire d'Ecologie Alpine (LECA) CNRS UMR 5553 2233, Université Joseph Fourrier, BP 53, 38041 Grenoble Cedex-9 and [2]INRIA Rhône-Alpes – Projet Bamboo, ZIRST-655 Avenue de l'Europe, 38334 Montbonnot Cedex, France

New resource when looking for primers:

Molecular Ecology Resources

WILEY

MOLECULAR ECOLOGY RESOURCES

RESOURCE ARTICLE OPEN ACCESS

The Ribosomal Operon Database: A Full-Length rDNA Operon Database Derived From Genome Assemblies

Anders K. Krabberød | Embla Stokke | Ella Thoen | Inger Skrede | Håvard Kauserud

Be aware of your primer biases!

Markers in DNA metabarcoding

- The ideal marker should:
  - Have primer sites that are shared by all target organisms
  - Be easy to amplify (suitable and matching melting temperatures)
  - 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!

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

The lenght of the marker must fit the sequencing technology!

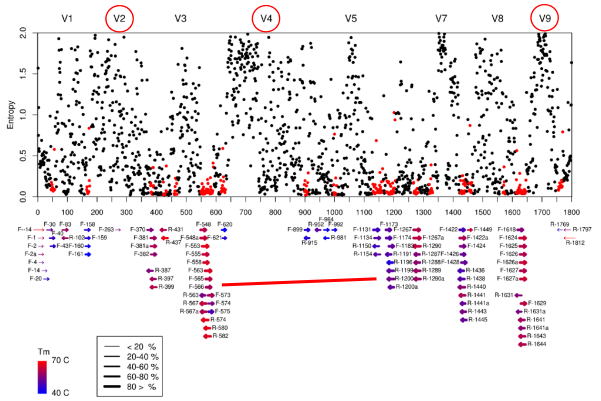


Figure 1. Nucleotide variability in 18S rRNA genes. Shannon entropy values of all eukaryotic alignment positions from the SILVA database along the 18S rRNA gene of *Saccharomyces cerevisiae*. Red dots mark consecutive nucleotide positions where at least 90% of ≥10 nt have entropy values lower than 0.2. 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

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!

Marker length variability

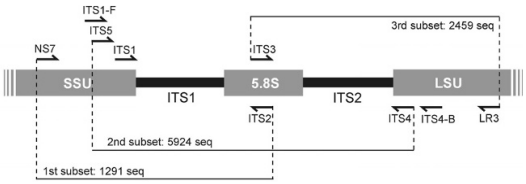
Bellemain et al. BMC Microbiology 2010, 10:189
http://www.biomedcentral.com/1471-2180/10/189

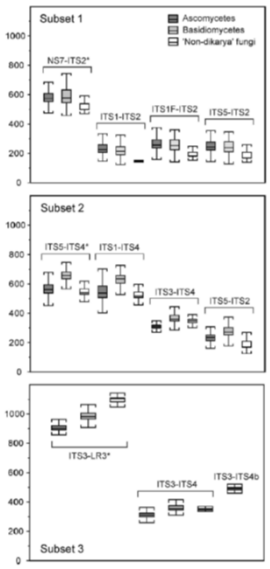
BMC Microbiology

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]





Some markers have a severe (taxonomic) length bias!

Marker length variability

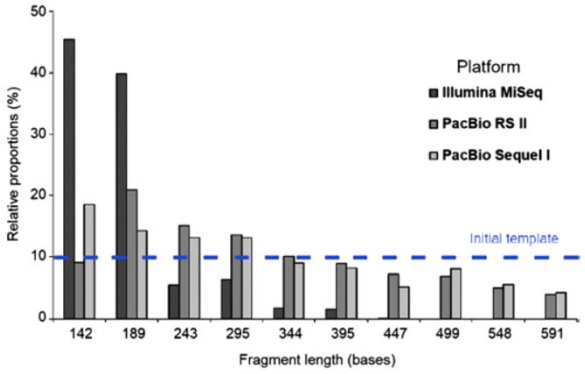
New Phytologist Research

Methods

Optimized metabarcoding with Pacific biosciences enables semi-quantitative analysis of fungal communities

Carles Castaño[1], Anna Berlin[1], Mikael Brandström Durling[1], Katharina Ihrmark[1], Björn D. Lindahl[2], Jan Stenlid[1], Karina E. Clemmensen[1]* and Ake Olson[1]*

Lenght biases introduced 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!

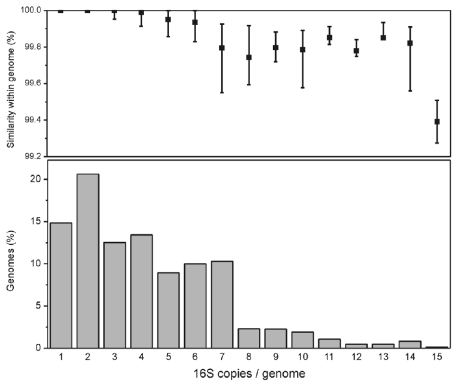
(Intra)genomic variability in 16S

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



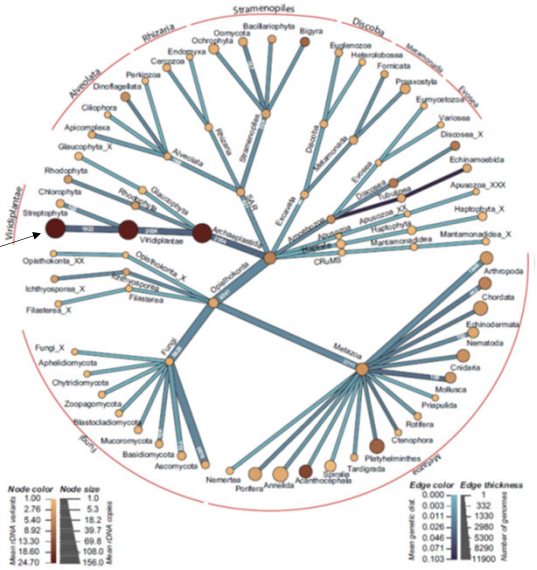
1,690 bacterial genomes with 7,081 16S copies analyzed

Inflates the diversity

(Intra)genomic variability in 16S

Inflates the diversity

High intragenomic variability (colour)



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!

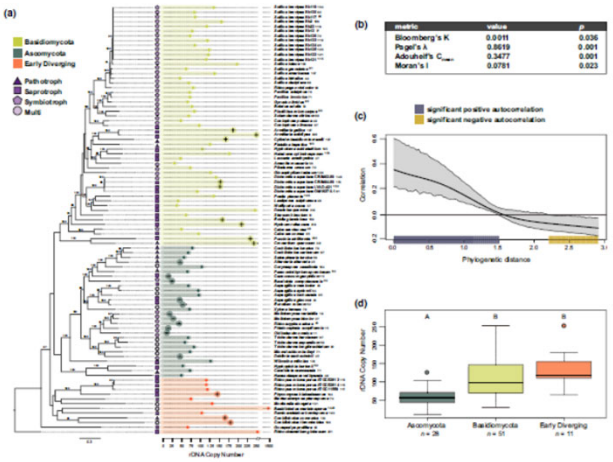
Copy number variation

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]



Introduces bias in semi-quantitative interpretations

Copy number variation

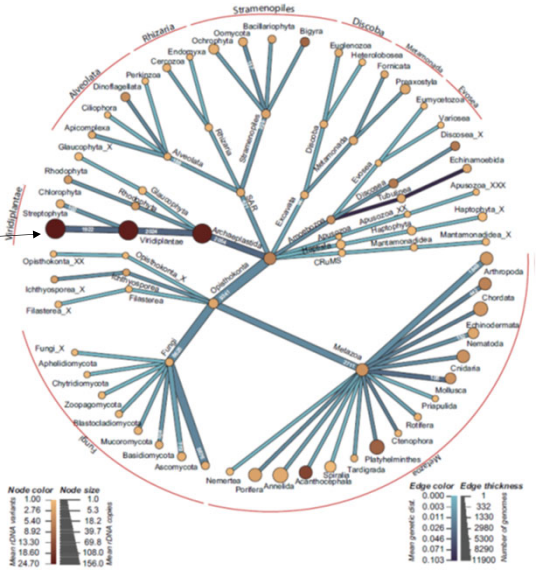
MOLECULAR ECOLOGY RESOURCES

RESOURCE ARTICLE OPEN ACCESS

The Ribosomal Operon Database: A Full-Length rDNA Operon Database Derived From Genome Assemblies

Anders K. Krabberød | Embla Stokke | Ella Thoen | Inger Skrede | Håvard Kauserud

High copy number (size)

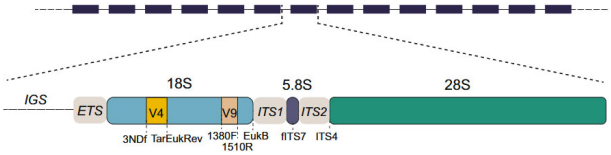


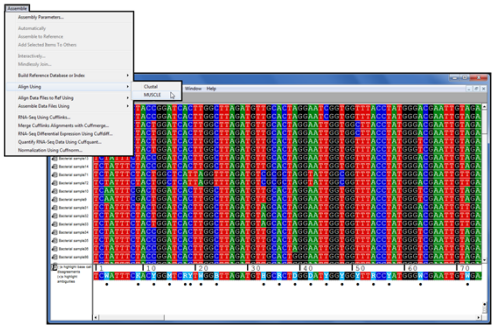
Introduces bias in semi-quantitative interpretations

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

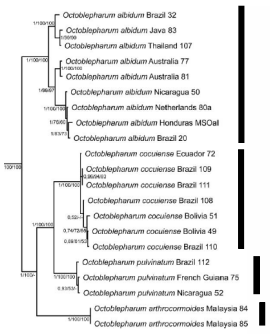


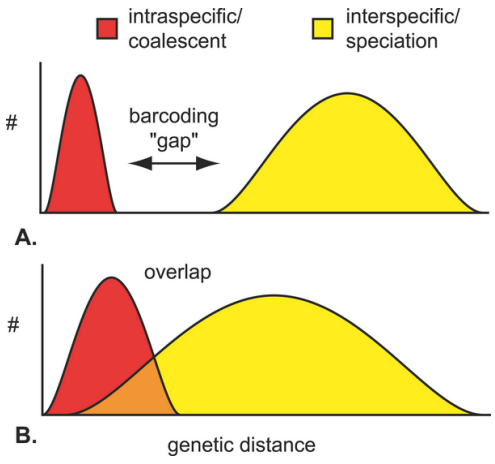


Markers in DNA metabarcoding

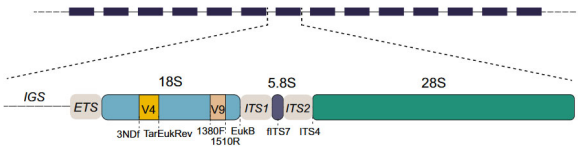
- The ideal marker should:
  - Have primer sites that are shared by all target organisms
  - Be easy to amplify
  - Be of appropriate length for efficient amplification and sequencing
  - Be of similar length
  - No intragenomic variation (i.e. no paralogs)
  - Be possible to align
  - Possess high interspecific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

Mind the barcoding gap





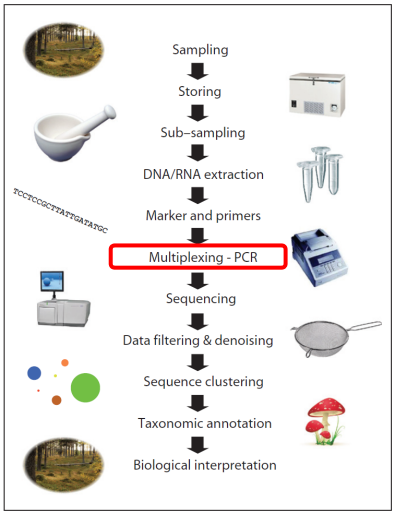
How variable is your marker?



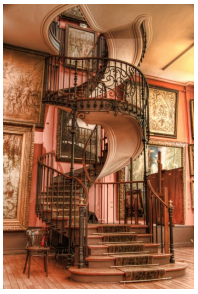
- 18S and 16S: Low variability, low intraspecific variation, low interspecific variation
- ITS: High variability, high intraspecific variation, high 'interspecific' variation

- Impact how the bioinformatics analyses should be conducted → there is no single way, no black box!

DNA metabarcoding - many steps



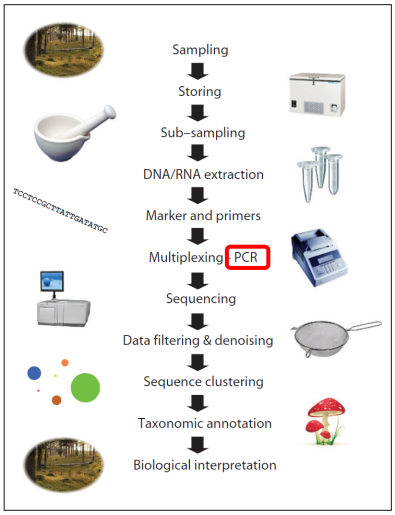
Lindahl et al. 2013



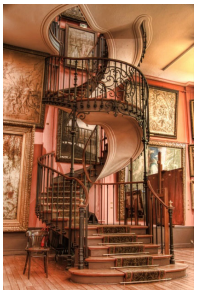
Many steps…

… to go wrong

DNA metabarcoding - many steps



Lindahl et al. 2013

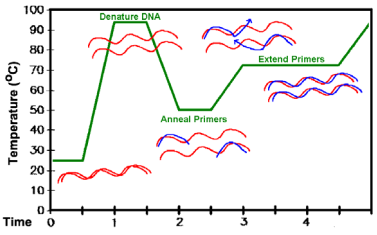


Many steps…

… to go wrong

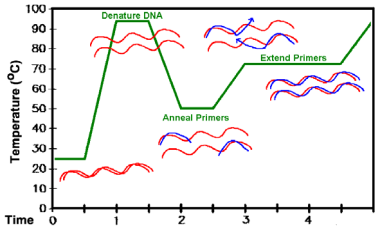
PCR

- 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!



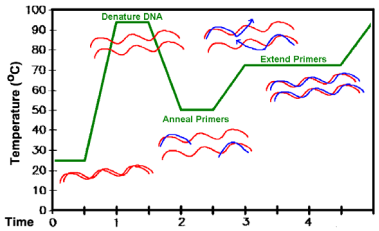
PCR

- 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!



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
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

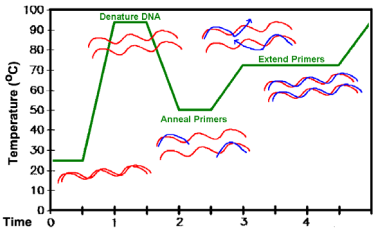
(Received November 19, 2012 / Accepted December 11, 2012)

. 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.

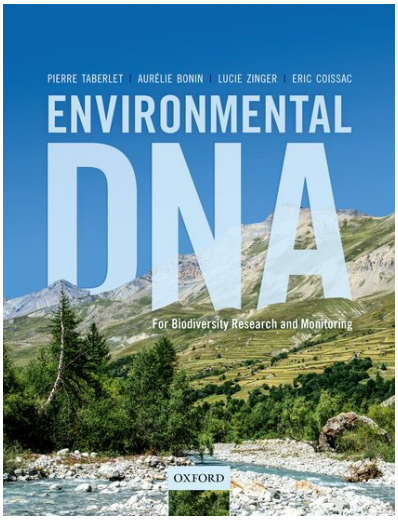
Keep the number of cycles as low as possible?!

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?



Taberlet et al: Multiple PCR amplifications and then remove outliers

Probably very context dependent: Low DNA concentrations and complex communities means that stochastisity during PCR is higher → The need for multiple amplicons is higher in such situations.

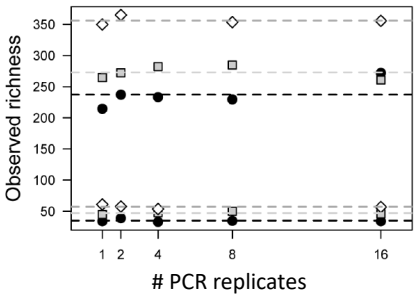
PCR replicates?

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



Sample types

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

PCR

Nested-PCR 😱

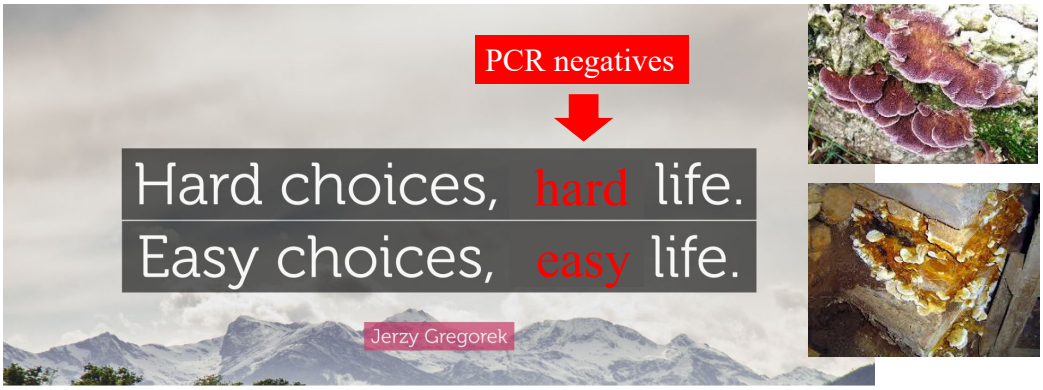
You should introduce PCR negatives!

PCR

PCR

You will very often get sequences in your extraction and PCR negatives!

PCR



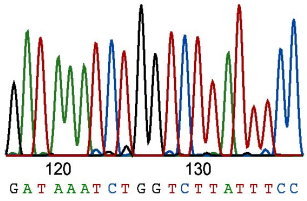
You will very often get sequences in your extraction and PCR negatives!

Sample types

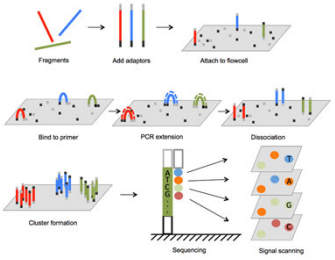
1. Biological replicates
2. Technical replicates
3. Extraction negatives
4. (PCR replicates)
5. PCR negatives

PCR induced errors

- PCR mutations: polymerase enzymes introduce erroneous nucleotides now and then, even those enzymes with proof-reading activity (like 1 in 1000 bp)
- Dependent on the technology whether these becomes «visible» or not:
  - 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

→ ITS amplicons from ~100 single spore cultures → sequencing

PCR-induced errors

Ecology and Evolution

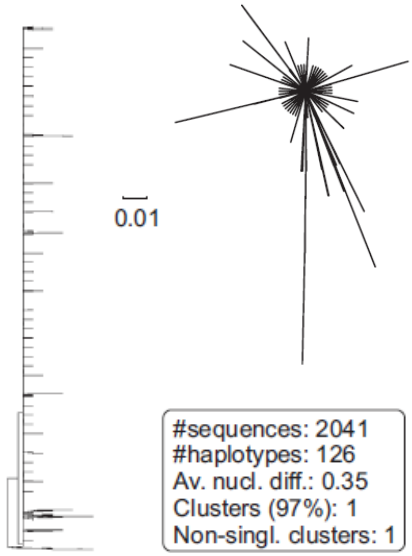
Open Access

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

(Could also argue that some of it are due to intragenomic varibility)



PCR-induced errors

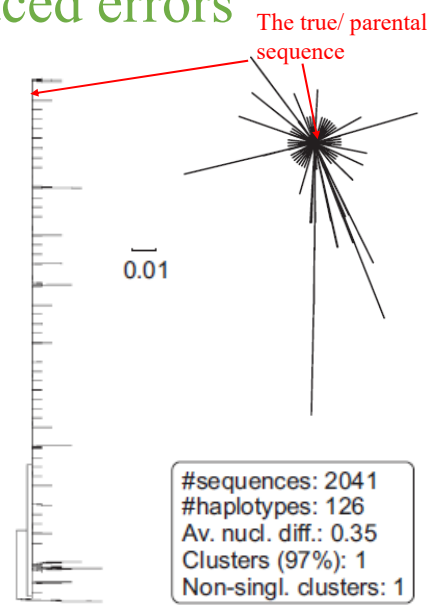
Ecology and Evolution

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

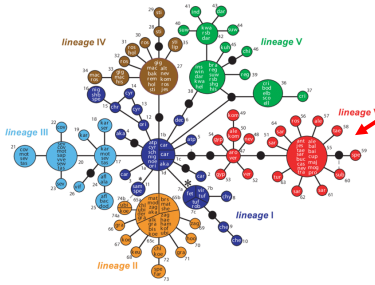
Ecology and Evolution

Open Access

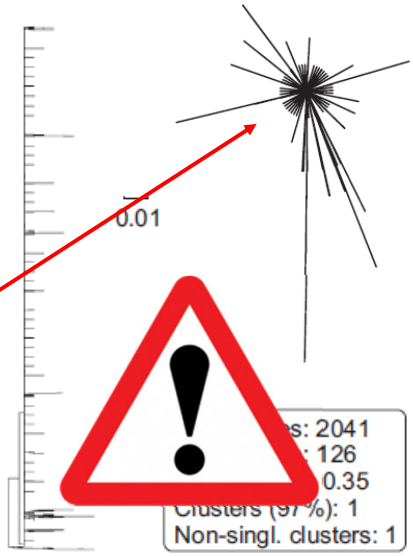
Employing 454 amplicon pyrosequencing to reveal intragenomic divergence in the internal transcribed spacer rDNA region in fungi

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

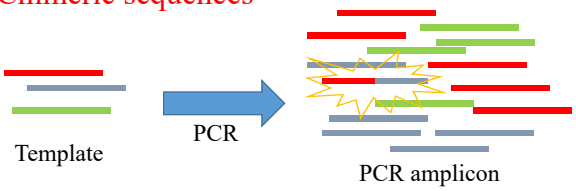


Same structure as natural variaton…



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

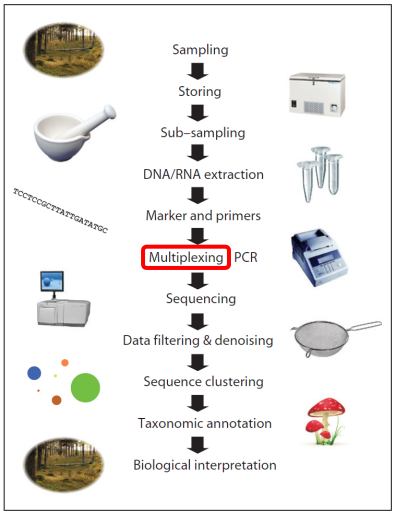
ANDERS BJØRNSGAARD AAS, MARIE LOUISE DAVEY and HÅVARD KAUSERUD
Section for Genetics and Evolutionary Biology (Evogene), Department of Biosciences, University of Oslo, P.O. Box 1066 Blindern, NO-0316 Oslo, Norway

Will inflate the diversity

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

DNA metabarcoding - many steps



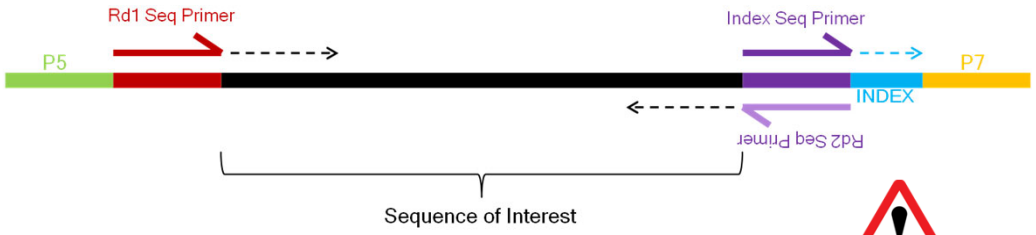
Lindahl et al. 2013

Different approaches to index the samples (as well as add on adaptors for sequencing):

- Directly during the PCR
- Nested PCR approach

Different ways to index samples and libraries

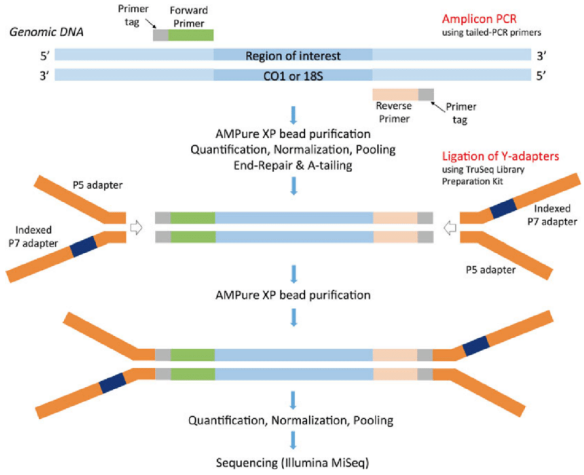
Fusion primer: One step process before sequencing



Only one index!

Different ways to index samples and libraries

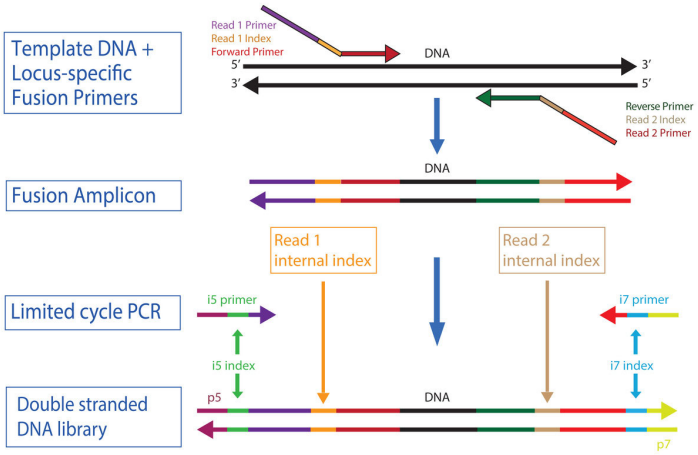
1 PCR + adaptor ligation: 2 steps before sequencing



«Our» standard approach

Different ways to index samples and libraries

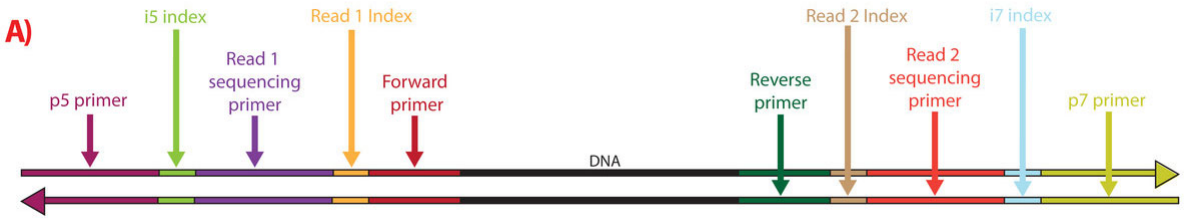
2 PCRs: 2 steps before sequencing



Nested PCR!

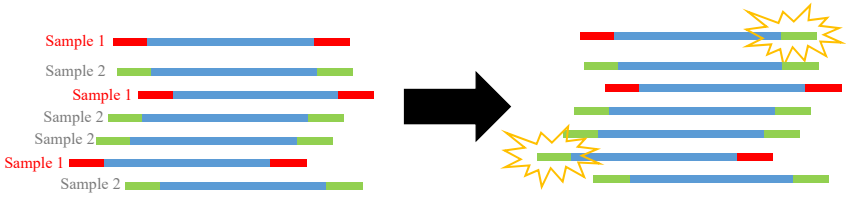
Different ways to index samples and libraries

2 PCRs: 2 steps before sequencing



A complex construct!

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



library sequences

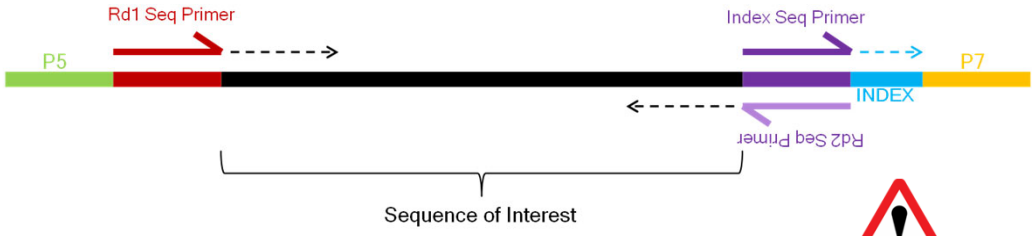
Samples

	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 if you don't tag properly in both ends!

Different ways to index samples and libraries

Fusion primer: One step process before sequencing



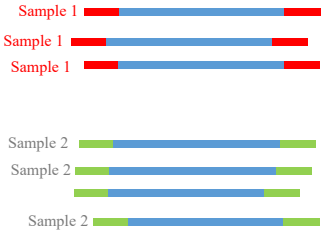
Only one index!

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

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

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

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

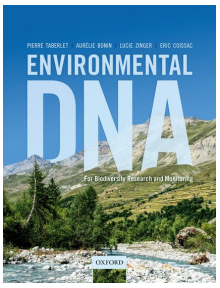
Received April 27, 2014; Revised January 29, 2015; Accepted January 30, 2015

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)



Sample types

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

Sample types

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

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 rxns

Multiple purposes!

- Tag switching
- Contamination
- Ability to delineate your species into «proper» OTUs

Sample types

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

Multiple purposes!
- Tag switching
- Contamination
- Ability to delineate your species into «proper» OTUs

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

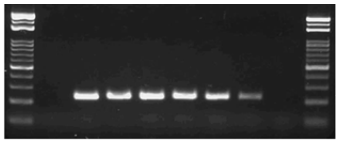
1. Biological replicates
2. Technical replicates
3. Extraction negatives
4. (PCR replicates)
5. PCR negatives
6. Positive control (mock community)

→ Only sequence if you have «positive» amplicons….? No!

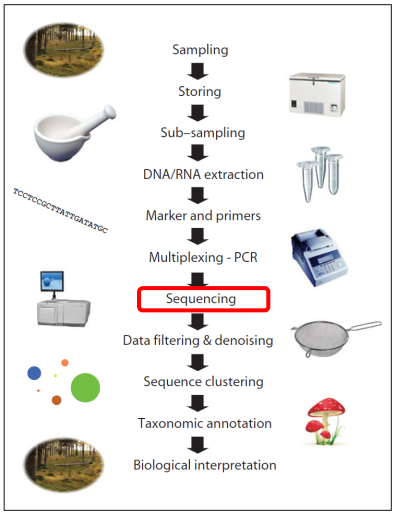
EDITORIAL — MOLECULAR ECOLOGY WILEY

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

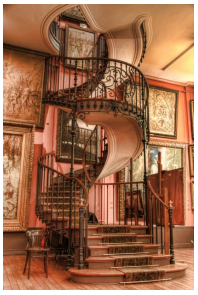
Zinger et al. 2019. Molecular Ecology Resources



DNA metabarcoding - many steps

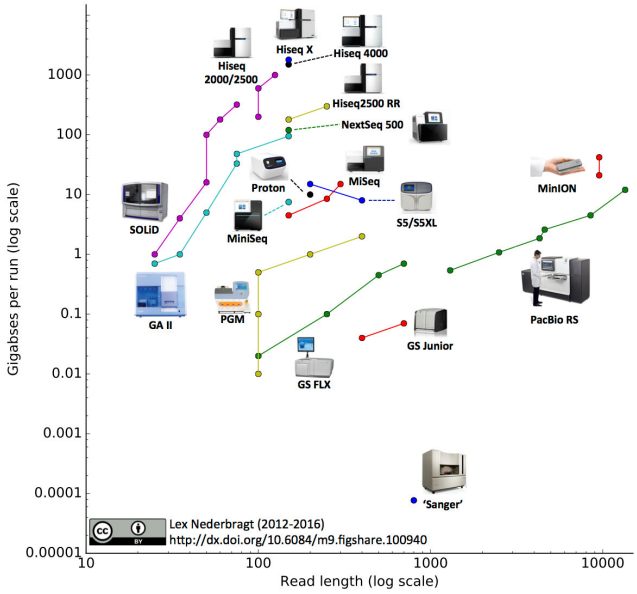


Lindahl et al. 2013

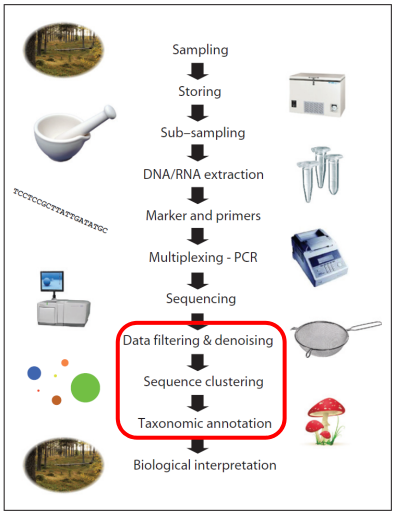


Many steps…

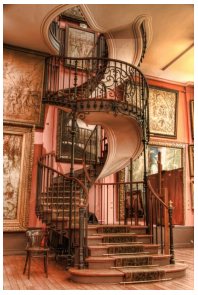
… to go wrong



DNA metabarcoding - many steps



Lindahl et al. 2013



Many steps…

… to go wrong

Bioinformatics – main steps

- QC Sequencing Results
- Demultiplexing
- Quality control, filtering, trimming
- Dereplication
- Denoising / OTU clustering
- Chimera removal
- OTU table construction
- Taxonomic assignment
- Removal of non-targets
- Normalization or rarefaction
- Downstream analysis and plotting

Bioinformatics – main steps

cutadapt

USEARCH

DADA2

PROTAX

VSEARCH

Qiime

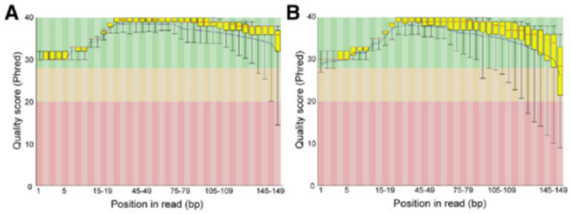
SINTAX

Obitools

mothur

Bioinformatics – main steps

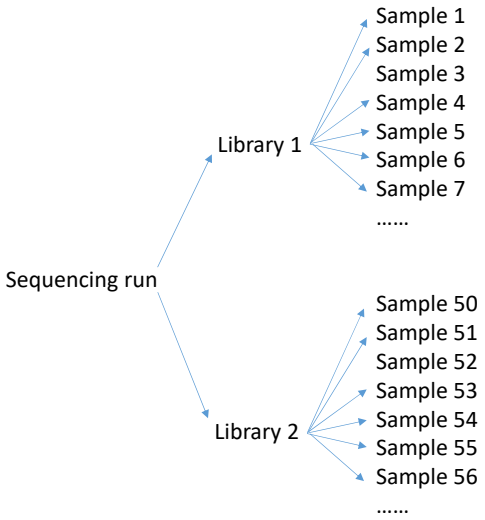
- QC Sequencing Results
- Demultiplexing
- Quality control, filtering, trimming
- Dereplication
- Denoising / OTU clustering
- Chimera removal
- OTU table construction
- Taxonomic assignment
- Removal of non-targets
- Normalization or rarefaction
- Downstream analysis and plotting



Remove
- Poor sequence quality
- Long/short sequences

Bioinformatics – main steps

- QC Sequencing Results
- Demultiplexing
- Quality control, filtering, trimming
- Dereplication
- Denoising / OTU clustering
- Chimera removal
- OTU table construction
- Taxonomic assignment
- Removal of non-targets
- Normalization or rarefaction
- Downstream analysis and plotting



Bioinformatics – main steps

- QC Sequencing Results
- Demultiplexing
- Quality control, filtering, trimming
- Dereplication
- Denoising / OTU clustering
- Chimera removal
- OTU table construction
- Taxonomic assignment
- Removal of non-targets
- Normalization or rarefaction
- Downstream analysis and plotting

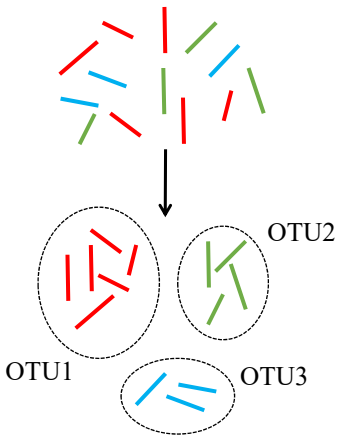
Bioinformatics – main steps

- QC Sequencing Results
- Demultiplexing
- Quality control, filtering, trimming
- Dereplication
- Denoising / OTU clustering
- Chimera removal
- OTU table construction
- Taxonomic assignment
- Removal of non-targets
- Normalization or rarefaction
- Downstream analysis and plotting

(k-mer based analyses)

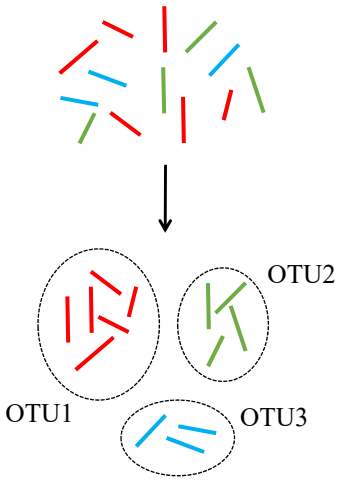
Bioinformatics – main steps

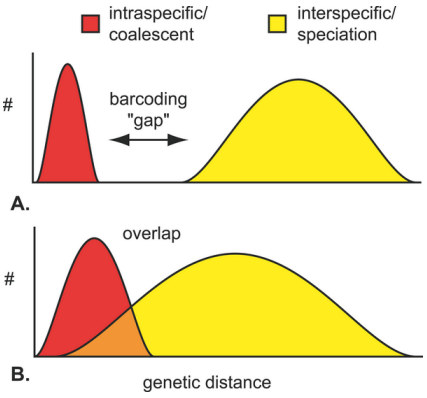
- QC Sequencing Results
- Demultiplexing
- Quality control, filtering, trimming
- Dereplication
- Denoising / OTU clustering
- Chimera removal
- OTU table construction
- Taxonomic assignment
- Removal of non-targets
- Normalization or rarefaction
- Downstream analysis and plotting



de novo versus closed (reference based) OTU construction?

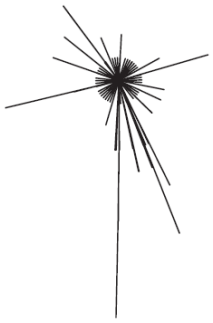
Denoising / OTU clustering

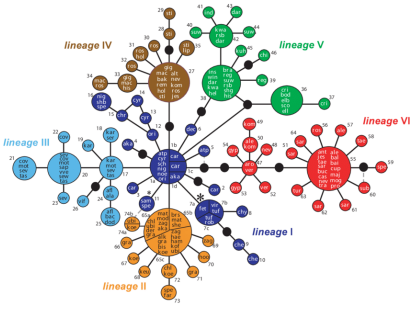




Requires a barcoding gap. How variable is your marker?

Denoising

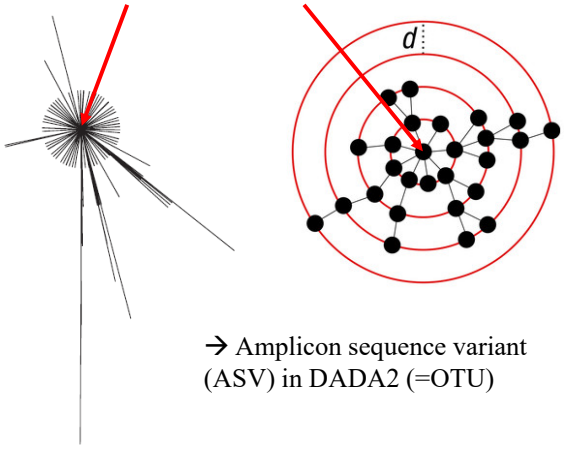




Important to consider both PCR induced errors versus natural sequence variation (intra- and interspecific)

Denoising

The parental («true») sequence

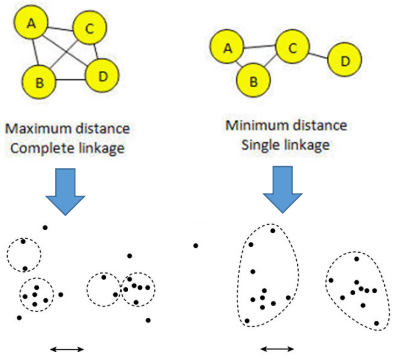


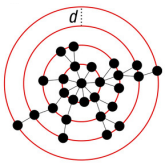
→ Amplicon sequence variant (ASV) in DADA2 (=OTU)

If you are working with a marker with no/limited intraspecific variability, like 18S or 16S, removing noise might be enough to construct your OTUs (or ASVs..). You don't need a second clustering step, then you will likely merge OTUs that should not be merged

OTU clustering

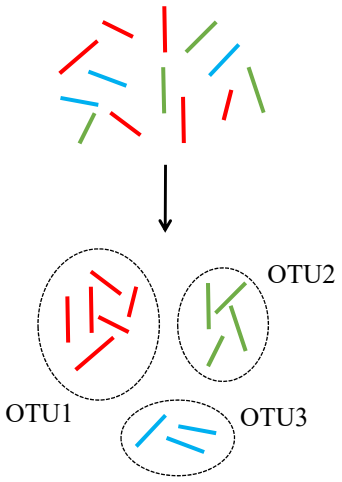
Many different clustering approaches to obtain the OTUs





SWARM

Denoising / OTU clustering

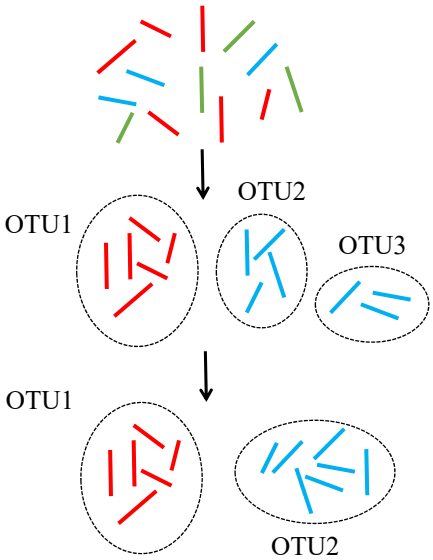


Different alternatives:

- Denoising only
- Denoising + clustering
- Clustering only

→ Depends on the marker you are using or computational possibilities/resources, etc.

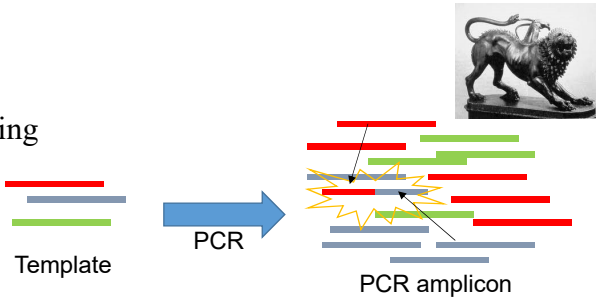
Tentative step: OTU merging



- Correct for over-splitting due to intraspecific variability (MUMU/LULU)

Bioinformatics – main steps

- QC Sequencing Results
- Demultiplexing
- Quality control, filtering, trimming
- Dereplication
- Denoising / OTU clustering
- Chimera removal
- OTU table construction
- Taxonomic assignment
- Removal of non-targets
- Normalization or rarefaction
- Downstream analysis and plotting

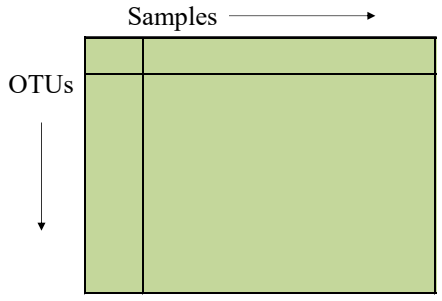


De novo versus reference based chimera checking

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

Bioinformatics – main steps

- QC Sequencing Results
- Demultiplexing
- Quality control, filtering, trimming
- Dereplication
- Denoising / OTU clustering
- Chimera removal
- OTU table construction
- Taxonomic assignment
- Removal of non-targets
- Normalization or rarefaction
- Downstream analysis and plotting



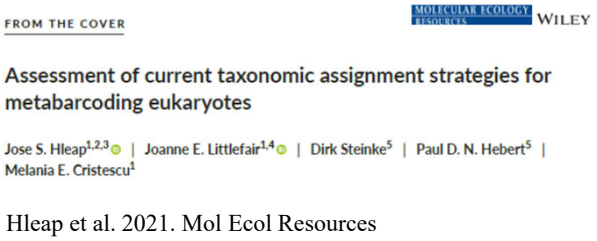
Bioinformatics – main steps

- QC Sequencing Results
- Demultiplexing
- Quality control, filtering, trimming
- Dereplication
- Denoising / OTU clustering
- Chimera removal
- OTU table construction
- Taxonomic assignment
- Removal of non-targets
- Normalization or rarefaction
- Downstream analysis and plotting

Simple matching (blast) → probabilistic assignment (e.g. protax)

Bioinformatics – main steps

- QC Sequencing Results
- Demultiplexing
- Quality control, filtering, trimming
- Dereplication
- Denoising / OTU clustering
- Chimera removal
- OTU table construction
- Taxonomic assignment
- Removal of non-targets
- Normalization or rarefaction
- Downstream analysis and plotting


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]

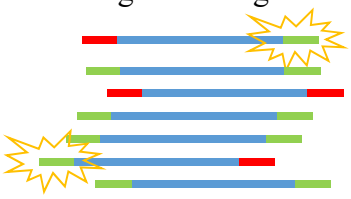
Hleap et al. 2021. Mol Ecol Resources

Bioinformatics – main steps

- QC Sequencing Results
- Demultiplexing
- Quality control, filtering, trimming
- Dereplication
- Denoising / OTU clustering
- Chimera removal
- OTU table construction
- Taxonomic assignment
- Removal of non-targets
- Normalization or rarefaction
- Downstream analysis and plotting

Non-specific amplification of other groups

Tag-switching



Contamination

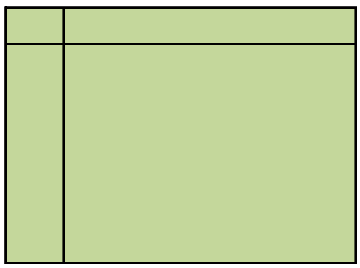
Should be very careful how you treat positive negatives!!

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

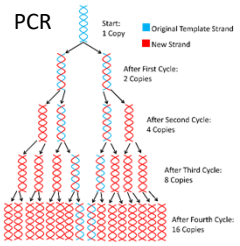
- QC Sequencing Results
- Demultiplexing
- Quality control, filtering, trimming
- Dereplication
- Denoising / OTU clustering
- Chimera removal
- OTU table construction
- Taxonomic assignment
- Removal of non-targets
- Normalization or rarefaction
- Downstream analysis and plotting



The sequencing depth typically varies a lot!

Bioinformatics – main steps

- QC Sequencing Results
- Demultiplexing
- Quality control, filtering, trimming
- Dereplication
- Denoising / OTU clustering
- Chimera removal
- OTU table construction
- Taxonomic assignment
- Removal of non-targets
- Normalization or rarefaction
- Downstream analysis and plotting



Be careful with resampling and transformations!

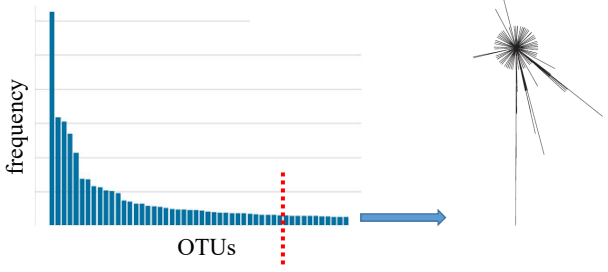
Check the effect from various data treatments options on the results!

Depends on the study aims!

Bioinformatics – main steps

- QC Sequencing Results
- Demultiplexing
- Quality control, filtering, trimming
- Dereplication
- Denoising / OTU clustering
- Chimera removal
- OTU table construction
- Taxonomic assignment
- Removal of non-targets
- Normalization or rarefaction
- Downstream analysis and plotting

Singleton removal?



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

Bioinformatics – main steps

- QC Sequencing Results
- Demultiplexing
- Quality control, filtering, trimming
- Dereplication
- Denoising / OTU clustering
- Chimera removal
- OTU table construction
- Taxonomic assignment
- Removal of non-targets
- Normalization or rarefaction
- Downstream analysis and plotting

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

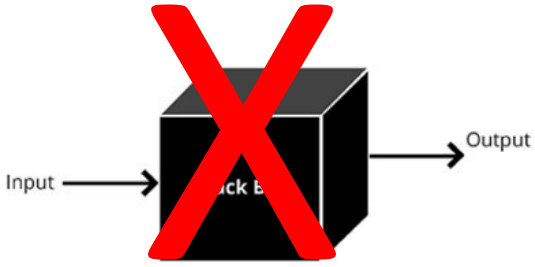
DNA metabarcoding–Need for robust experimental designs to draw sound ecological conclusions

Zinger et al. 2019, Molecular Ecology Resources

Secondary phase
- Established scientific approach with a set of widely accepted guidelines

Take home message

Which methods to use? → No general answer – it is context dependent



Learning goals

- Have a conceptual understanding of the various steps during bioinformatics analyses of metabarcoding data .
- Understand and argue for why you make your choices and use various bioinformatics approaches in different situations. Learn to be critical towards DNA metabarcoding data and how to evaluate them.
- Develop basic knowlege in important programs, like cutadapt, DADA2, SWARM and VSEARCH.
- Obtain knowledge about long-read metabarcoding and phylogenetic placement.
- Obtain insight in possible downstream analyses.
- By discussing with your fellow students – learn about different DNA metabarcoding projects.

Report

From «wild west» towards an established approach

MOLECULAR ECOLOGY WILEY

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

Zinger et al. 2019, Molecular Ecology Resources

International Microbiology (2024) 27:719–730
https://doi.org/10.1007/s10123-023-00422-5

RESEARCH

A new spike-in-based method for quantitative metabarcoding of soil fungi and bacteria

Miguel Camacho-Sanchez

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2024, VOL. 00, NO. 0, 1–15: Applications and Case Studies
https://doi.org/10.1080/01621459.2024.2412362

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

eDNAPlus: A Unifying Modeling Framework for DNA-based Biodiversity Monitoring

Alex Diana[a], Eleni Matechou[b], Jim Griffin[c], Douglas W. Yu[d,e], Mingjie Luo[f], Marie Tosa[g], Alex Bush[h], and Richard A. Griffiths[i]

- Primary phase
- Secondary phase

- Tertiary phase
  - Improved taxonomic resolution
  - Quantitative metabarcoding
  - Quantification of detection probability