

BIO9905MERG1 – Bioinformatics for Environmental Sequencing (DNA metabarcoding)

Welcome!



BIO9905MERG1 – Bioinformatics for Environmental Sequencing (DNA metabarcoding)



Organization

- Mix of lectures + discussions + hands on sessions
- Zoom: Stay muted (when not asking questions). Please use your camera, if you are comfortable with that.
- Basic → advanced. High variation in experience. Experienced students help less experienced → mattermost
- Mattermost - our channel for written communication / chatting during the course. Inform if you have not been invited: haavarka@ibv.uio.no
- Group work: Please engage in this, otherwise it will not be a success ☺ Group work will happen in zoom break out rooms. Group overview is available at github
- Hands-on sessions in R. Some assistance in installing packages is provided Monday 4pm.
- Report: Instructions provided at: https://github.com/krabberod/BIO9905MERG1_V21/
- Learning by doing, please have some patience with us..



Mattermost

The screenshot shows the Mattermost web interface for the 'Town Square' channel. The channel has a guest message board. The message list shows numerous guests joining the channel from various IP addresses and times, including @srijana, @david_williams, @carolinewmkartoon, @eliseek, @hegevd, @bona, @moe.saran, and @ahrendra. The interface includes a sidebar with channels like 'Course organizers', 'Message to course admin', 'Off topic', and 'Town Square'. A search bar and a help link are also visible.

Organization

- Mix of lectures + discussions + hands on sessions
- Zoom: Stay muted (when not asking questions). Please use your camera, if you are comfortable with that.
- Basic → advanced. High variation in experience. Experienced students help less experienced → mattermost
- Mattermost - our channel for written communication / chatting during the course. Inform if you have not been invited: haavarka@ibv.uio.no
- Group work: Please engage in this, otherwise it will not be a success ☺ Group work will happen in zoom break out rooms. Group overview is available at github
- Hands-on sessions in R. Some assistance in installing packages is provided Monday 4pm.
- Report: Instructions provided at:
https://github.com/krabberod/BIO9905MERG1_V21/
- Learning by doing, please have some patience with us..



Registration

<https://nettskjema.no/a/201061>

Introduction to DNA metabarcoding

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

INVITED REVIEWS AND SYNTHESSES

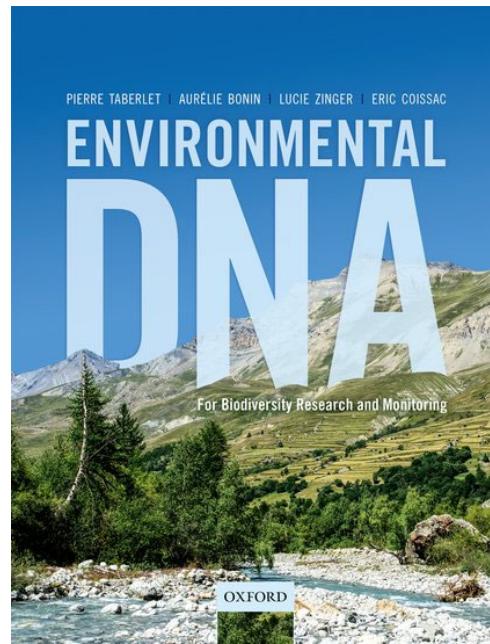
WILEY MOLECULAR ECOLOGY

Environmental DNA metabarcoding: Transforming how we survey animal and plant communities

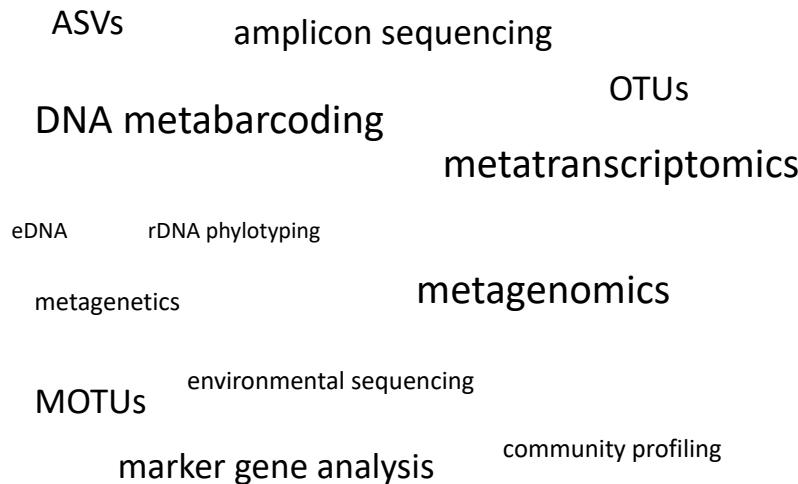
Kristy Deiner¹ | Holly M. Bik² | Elvira Mächler^{3,4} | Mathew Seymour⁵ |
 Anais Lacoursière-Roussel⁶ | Florian Altermatt^{3,4} | Simon Creer⁵ | Iliana Bista^{5,7} |
 David M. Lodge¹ | Natasha de Vere^{8,9} | Michael E. Pfrender¹⁰ | Louis Bernatchez⁶



Pierre Taberlet



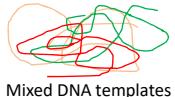
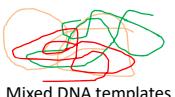
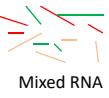
Some confusing / important terms



Some important terms

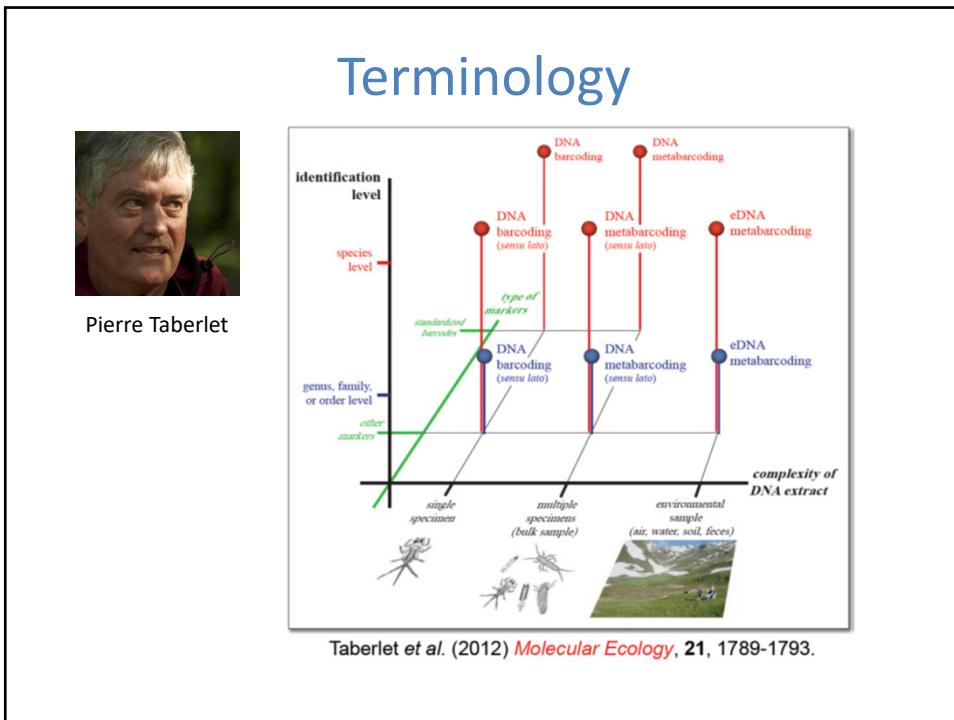
- DNA barcoding  → Sequence variation in a single locus (e.g. ITS) in a single specimen
- Metabarcoding  → Sequence variation in a single locus (e.g. ITS) in a community
Mixed DNA templates
- Metagenomics  → Genome wide sequence variation in a community
Mixed DNA templates
- Metatranscriptomics  → cDNA sequence variation in a community
Mixed RNA

Some important terms

- Metabarcoding  Sequence variation in a single locus (e.g. 16S) **Who are there?**
- Metagenomics  Genome wide sequence variation **Which genes (and who) are there?**
- Metatranscriptomics  cDNA sequence variation **Who are active and doing what?**

Some confusing / important terms

ASVs	amplicon sequencing	OTUs
DNA metabarcoding		
eDNA	rDNA phylotyping	metatranscriptomics
metagenetics		
MOTUs	environmental sequencing	metagenomics
marker gene analysis		
community profiling		



DNA metabarcoding

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



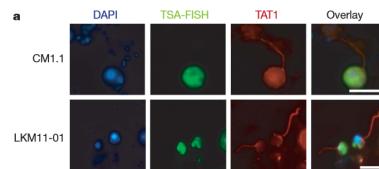
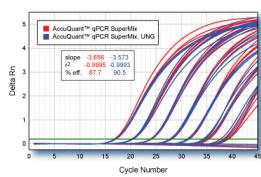
DNA metabarcoding

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



DNA metabarcoding

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



DNA metabarcoding: From ‘wild west’ towards an established approach?

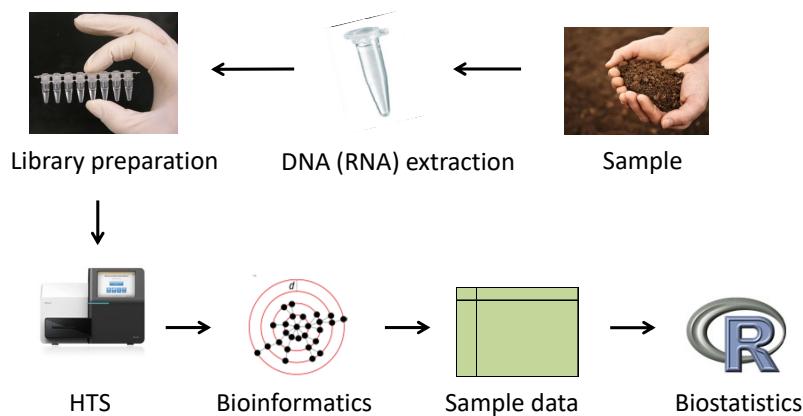


- Primary phase**
- Poor replication
 - Lack of controls
 - Lack of insight into important biases
 - Poor bioinformatics approaches

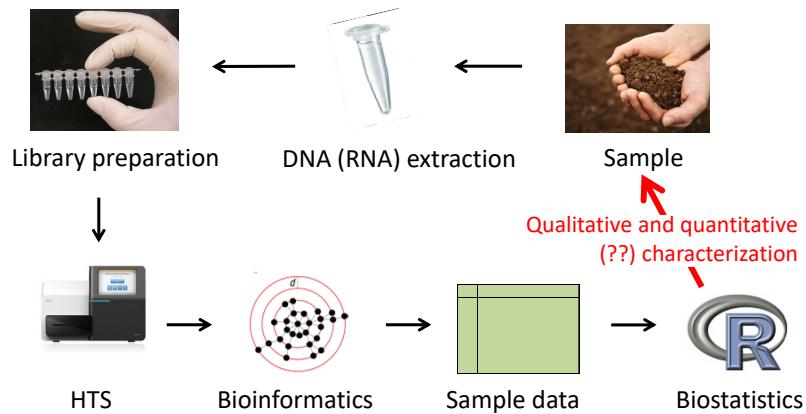
EDITORIAL MOLECULAR ECOLOGY WILEY
DNA metabarcoding—Need for robust experimental designs to draw sound ecological conclusions
Zinger et al. 2019, Molecular Ecology Resources

Established scientific approach with a set of widely accepted guidelines

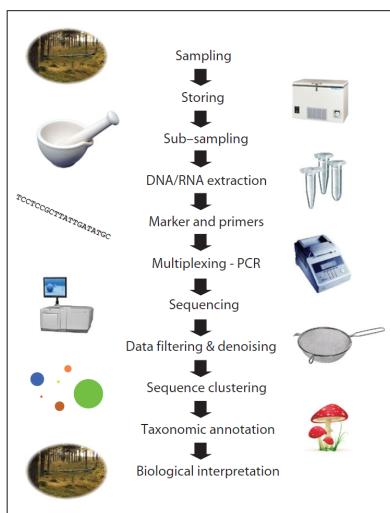
General workflow in DNA metabarcoding studies



General workflow in DNA metabarcoding studies



DNA metabarcoding - many steps



Many steps...



... to go wrong

Lindahl et al. 2013

EDITORIAL

MOLECULAR ECOLOGY WILEY

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

Zinger et al. 2019. Molecular Ecology Resources

RESEARCH ARTICLE

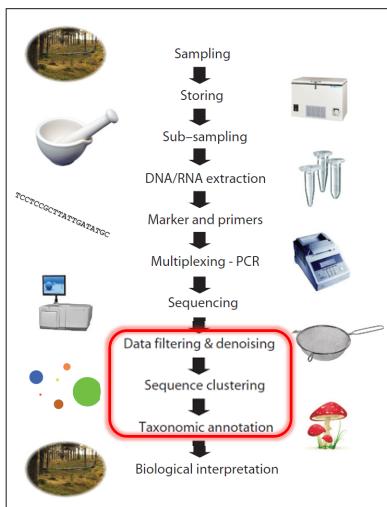
Methods in Ecology and Evolution 

Scrubinizing key steps for reliable metabarcoding of environmental samples

Antton Alberdi¹  | Ostaizka Aizpurua¹ | M. Thomas P. Gilbert^{1,2,3} | Kristine Bohmann^{1,4}

Alberdi et al. 2017

DNA metabarcoding - many steps



```

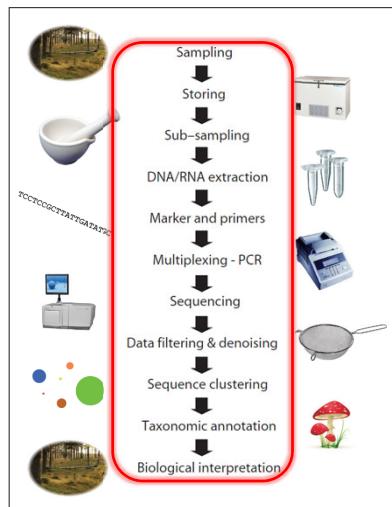
graph TD
    Sampling --> Storing
    Storing --> Subsampling
    Subsampling --> DNAExtraction[DNA/RNA extraction]
    DNAExtraction --> MarkerPrimers[Marker and primers]
    MarkerPrimers --> MultiplexingPCR[Multiplexing - PCR]
    MultiplexingPCR --> Sequencing
    Sequencing --> DataFiltering[Data filtering & denoising]
    DataFiltering --> SequenceClustering[Sequence clustering]
    SequenceClustering --> TaxonomicAnnotation[Taxonomic annotation]
    TaxonomicAnnotation --> BioInterpretation[Biological interpretation]
    
```

Lindahl et al. 2013

Many steps...

... to go wrong

DNA metabarcoding - many steps



Lindahl et al. 2013

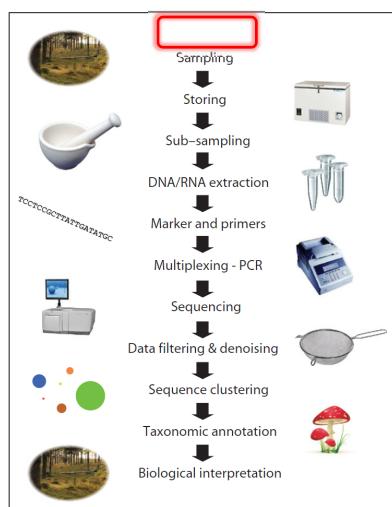


Many steps...



... to go wrong

DNA metabarcoding - many steps



Lindahl et al. 2013



Many steps...



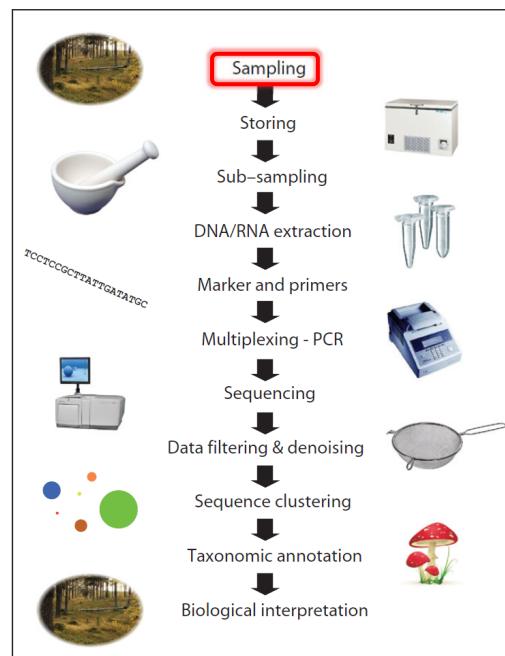
... to go wrong

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
(which you might not know anything about..).



Representativeness (in space)

- Many communities highly heterogenous
- Should obtain samples that are representative
- 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?
- When comparing across different areas: May use phenological markers to obtain the 'same' temporal phase.



«Replicate or lie»

Opinion
Replicate or lie

James J. Prosser
Institute of Biological and Environmental Sciences,
University of Aberdeen, Crookes Building, St.
Marie's Street, Aberdeen, AB24 2UE, UK

Introduction
Andren and colleagues (2005) recently published a paper in this journal that has been widely cited and used during our early years of research, but has also been heavily criticized. The paper was highly lauded for the majority of measured ecogeomes have been found to be significantly different from each other. In this problem, and my authority of the consequences, are the authors guilty of replicating their results? I believe so.

The solution is to conduct more replicates with strict adherence to the methods used. This is not always possible, given the nature of the samples and the environment. The use of sequencing techniques such as pyrosequencing and metagenomics can go beyond these techniques and benefit studies of microdiversity.

Why replicate?
The main reason the journal could be filled with articles describing and discussing applications of statistical analysis in environmental microbiology is the lack of a basic and fundamental aspect – the need for replication. In general, we cannot accept the results of a single observation, without trying to conceive benthic abundance in two sites. We can only accept the results of two observations with one site, giving weight 1×10^0 and 3.2×10^{-1} . One observation is not enough, because it is greater than one site for the other.

Many researchers have argued that the studies in their basic methods reduce and/or variability arising from

Received 2 December 2006, accepted 2 January 2007. "For correspondence": James J. Prosser, Institute of Biological and Environmental Sciences, University of Aberdeen, St. Marie's Street, Aberdeen, AB24 2UE, UK.
© 2010 Society for Applied Microbiology and Blackwell Publishing Ltd

Clone library analysis and pyrosequencing

	Number of articles	% with replicates
<i>Appl Environ Microbiol</i>	60	23
<i>Environ Microbiol</i>	47	15
<i>FEMS Microbiol Ecol</i>	29	24
<i>ISME J</i>	23	13
<i>Microbial Ecol</i>	22	9
Total	181	18

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

Prosser JJ. 2010, Environmental Microbiology

Received: 4 October 2015 | Revised: 10 May 2016 | Accepted: 14 May 2016
DOI: 10.1111/1365-2788.12907

INVITED TECHNICAL REVIEW

Towards robust and repeatable sampling methods in eDNA-based studies

Ian A. Dickie^{1,2} | Stephane Boyer^{3,4} |
Paul P. Gardner² | Ian D. Hogg^{7,8} | Robe
Andreas Makiola¹ | Sergio E. Morales¹¹ |

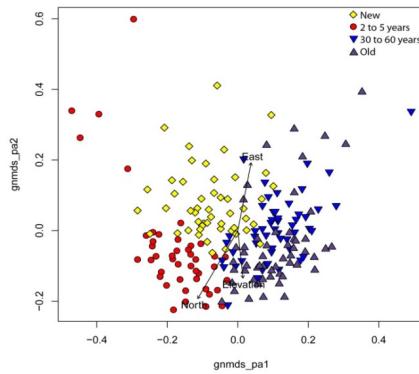
Abstract

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 bioinformatics 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

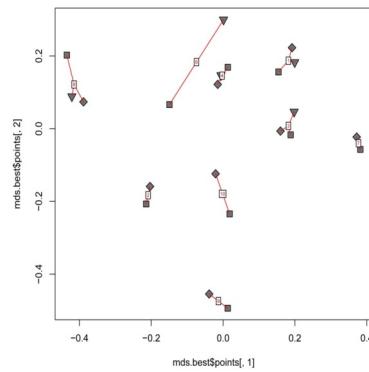


Davey *et al.* 2014. FEMS Microbial Ecology

Technical replicates

- Some samples should be analyzed multiple times
- Reveal the variability (experimental error) of the analysis technique → allows to set limits for what is meaningful and significant data

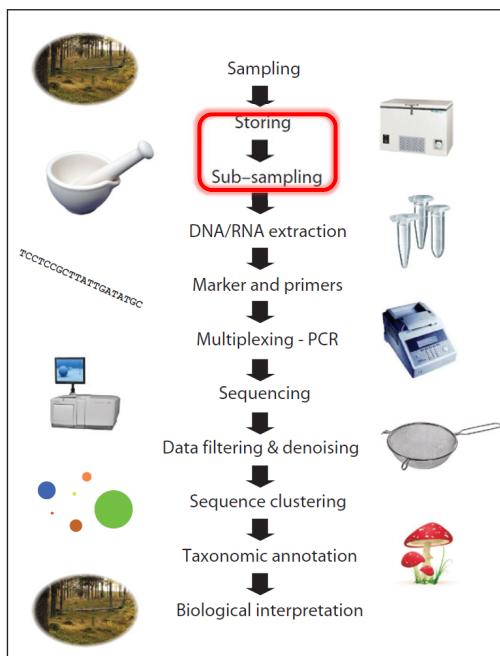
Communities with low DNA content!!



Davey *et al.* 2014. FEMS Microbial Ecology

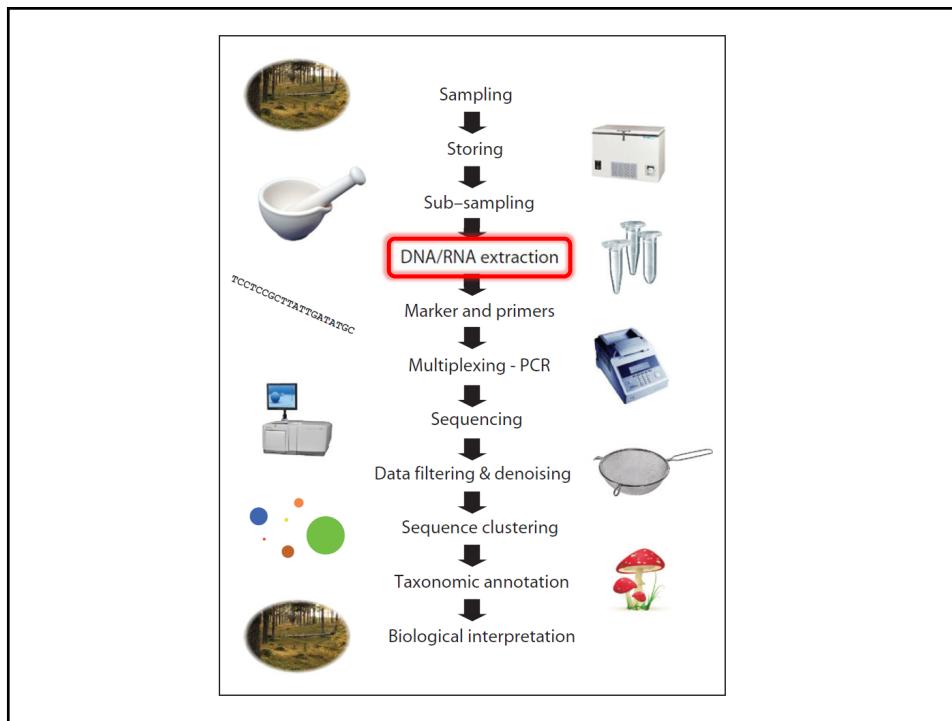
Sample types

1. Biological replicates
2. Technical replicates



Storing

- Unappropriate 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



DNA extraction

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

S.M. Dineen^{1,2}, R. Aranda IV^{1,2}, D.L. Anders³ and J.M. Robertson²

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

NOTE / NOTE

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

Lihui Xu, Sabine Ravansky, John Larson, and Higgins Nicollaisen

Molecular biology, genetics and biotechnology

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

Katerina Fliegerova^{a,1}, Ilma Tapiö^b, Aurelie Bonin^a, Jakub Młazek^a, Maria Luisa Callegari^c, Paolo Bini^c, Alireza Bayat^d, Johanna Vilkkilä^d, Jan Kopečný^a, Kevin J. Shingfield^{a,1}, Frederic Boyer^c, Eric Coissac^c, Pierre Taberlet^c, R. John Wallace^a

^a Institut für Biologische und Biochemische Analysen (IBBA), Leibniz Institute for Agricultural and Environmental Hygiene, University of Bayreuth, Bayreuth, Germany, ^b Department of Animal Health and Welfare, University of Helsinki, Helsinki, Finland, ^c INRAE, UMR AgroBiosphère, Paris, France, ^d Department of Animal Health and Welfare, University of Helsinki, Helsinki, Finland

PLOS ONE

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

Gemma Henderson¹, Faith Cox¹, Sandra Kittelmann¹, Vanidéh Heidarian Miri², Michael Zethor¹, Samantha J. Noël¹, Garry C. Waghorn², Peter H. Jansson^{1*}

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

Nicholas A. Kennedy¹, Alan W. Walker², Susan H. Berry³, Sylvia H. Duncan⁴, Freda M. Farquharson⁴, Petra Louis⁴, John M. Thomson⁵, UK IBD Genetics Consortium, Jack Satsangi¹, Harry J. Flint⁴, Julian Parkhill², Charlie W. Lee^{1*}, Georgina L. Hold^{3,*}

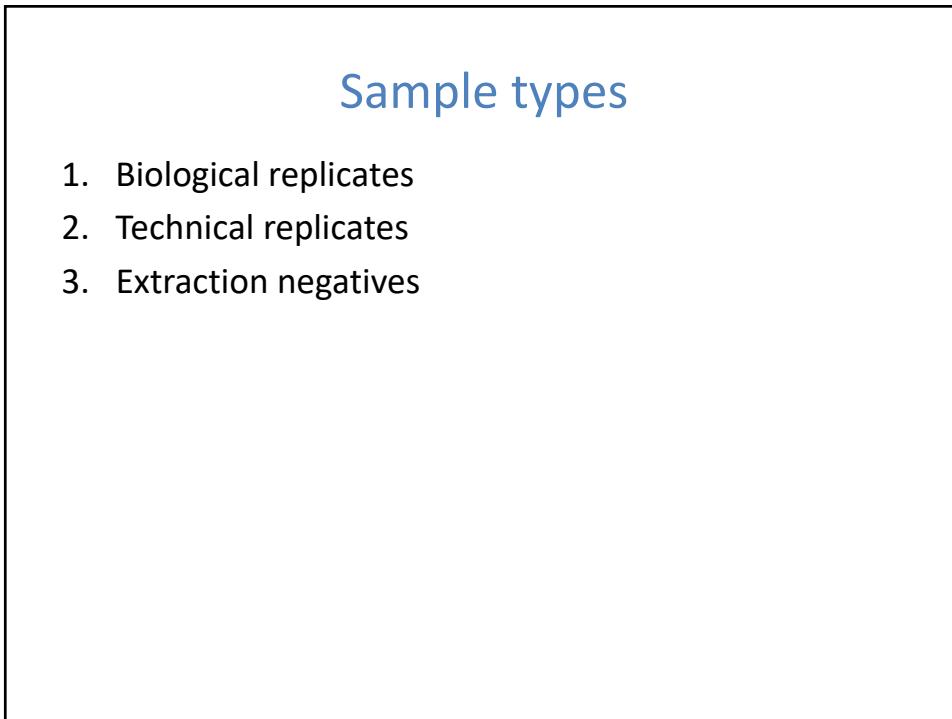
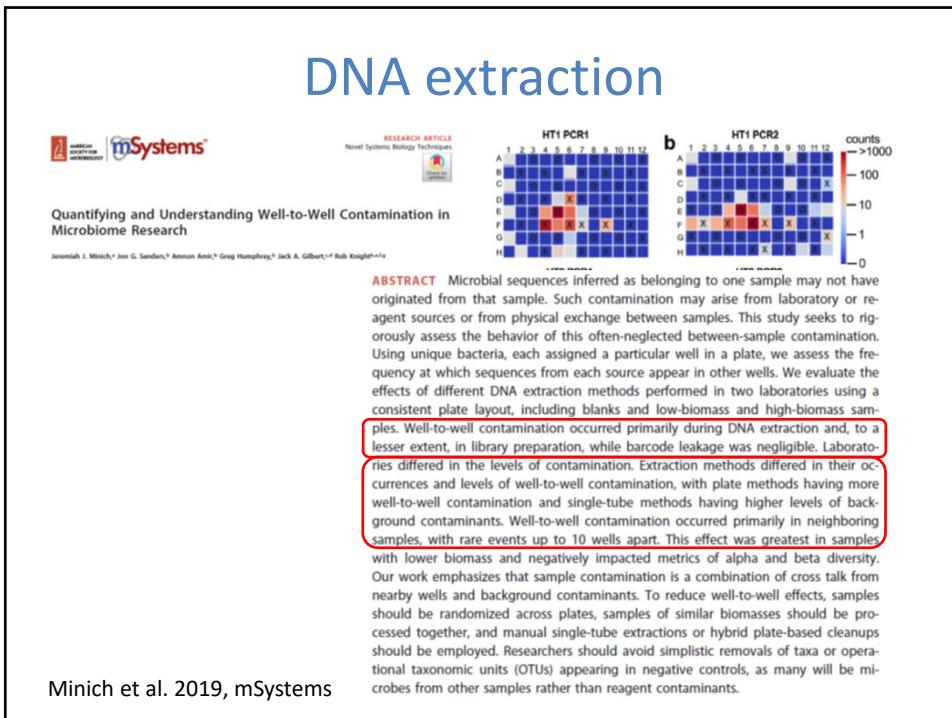
DNA extraction

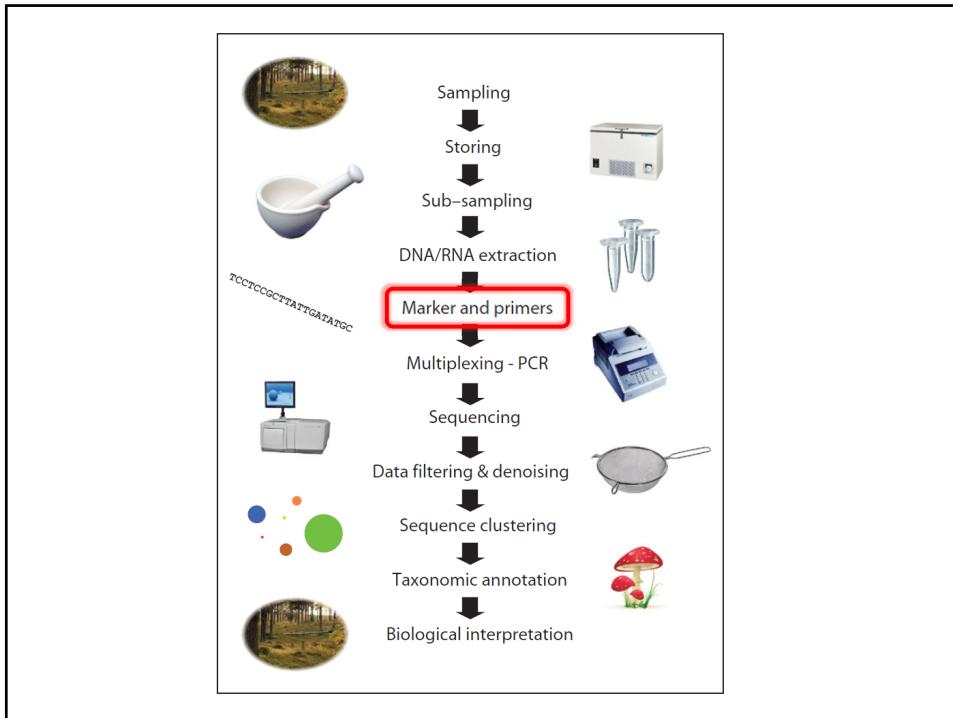
- Should yield high and uniform amounts of DNA
- Concentration of PCR inhibitors minimized
- Same protocol for all samples!
- If no proper literature are available on your study system → conduct a pilot?!
- Extraction negatives!





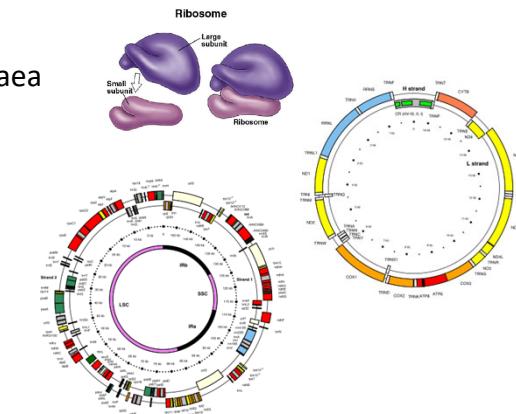
- MoBio Power Soil?
- FastDNA kit for Soil?
- EZNA Soil kit?
- CTAB + cleanup kit?





Markers used in DNA metabarcoding

- Standard markers (>500 bp):
 - 18S: Eukaryotes
 - 16S: Bacteria/archaea
 - ITS: Fungi & plants
 - COI: Metazoa
 - *RbcL*: Plants
 - *trnL*: 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)
 - 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)
 - Be possible to align
 - Have high interspecific variation
 - Have low intraspecific variation
- No known markers meet all these requirements!

ITS and primer bias

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

RESEARCH ARTICLE **Open Access**

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

Eva Bellemain¹, Tor Carlsen², Christian Brochmann¹, Eric Coissac³, Pierre Taberlet³ and Håvard Kauserud²

Marker Pair	1st subset (%)	2nd subset (%)	3rd subset (%)
ITS1-ITS2	~38	~95	~98
ITS2-ITS2	~82	~62	~98
ITS1-ITS4	~70	~75	~92
ITS2-ITS4	~80	~90	~92

Bellemain et al. 2010, BMC Microbiology

Markers in DNA metabarcoding

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

New Phytologist

Research

Methods

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

Carles Castaño¹, Anna Berlin¹, Mikael Brandström Durling¹, Katharina Ihrmark¹, Björn D. Lindahl², Jan Stenlid¹, Karina E. Clemmensen^{1*} and Ake Olson^{1*}

Fragment length (bases)	Illumina MiSeq (%)	PacBio RS II (%)	PacBio Sequel I (%)
142	~45	~10	~18
189	~40	~20	~15
243	~5	~15	~15
295	~5	~15	~15
344	~5	~10	~10
395	~2	~8	~8
447	~2	~5	~5
499	~2	~7	~7
548	~2	~5	~5
591	~2	~4	~4

Markers in DNA metabarcoding

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

(Intra)genomic variability in 16S

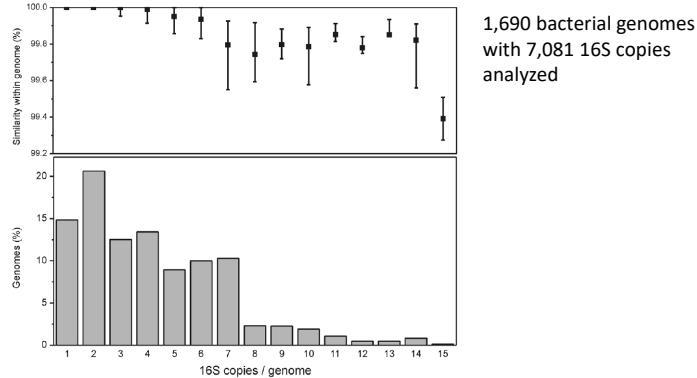
OPEN  ACCESS Freely available online

 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



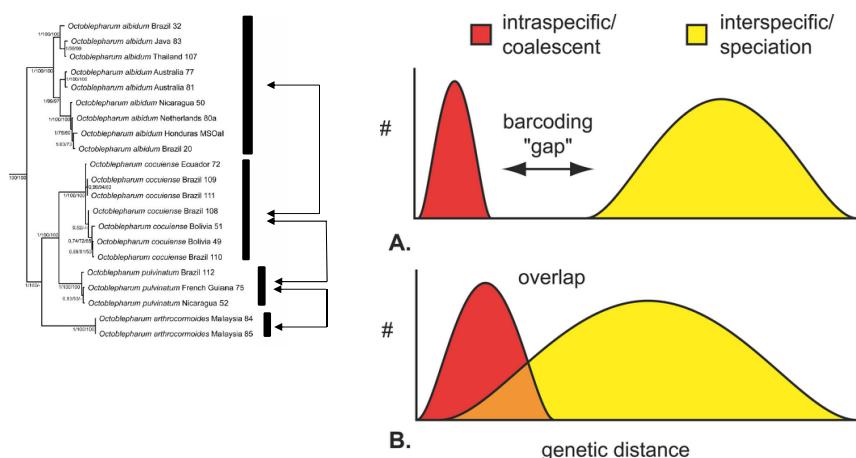
Markers in DNA metabarcoding

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

Markers in DNA metabarcoding

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

The barcoding gap



How conserved/variable are the marker?

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



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

Multiple markers/primers?

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



RESEARCH

Open Access



Evaluating a multigene environmental DNA approach for biodiversity assessment

Alexei J. Drummond^{1,2*}, Richard D. Newcomb^{1,3,4}, Thomas R. Buckley^{1,3,5}, Dong Xie^{1,2}, Andrew Dopheid^{1,3,4}, Benjamin CM Potter^{1,3}, Joseph Heled^{1,2}, Howard A. Ross^{1,3}, Leah Tooman^{1,4}, Stefanie Grosser^{1,5}, Duckchul Park⁵, Nicholas J. Demetras⁸, Mark I. Stevens^{6,7}, James C. Russell^{1,3,9}, Sandra H. Anderson³, Anna Carter^{1,10} and Nicola Nelson^{1,10}

- Often lack proper reference databases for multiple markers..

Long-read metabarcoding

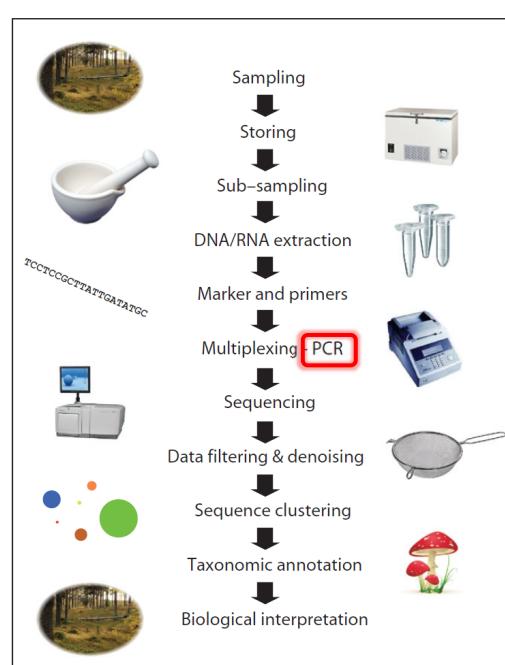
RESOURCE ARTICLE

MOLECULAR ECOLOGY
RESOURCES

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

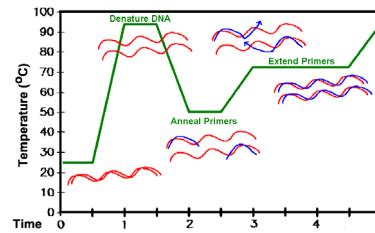
Mahwash Jamy¹ | Rachel Foster² | Pierre Barbera³ | Lucas Czech³ | Alexey Kozlov³ | Alexandros Stamatakis^{3,4} | Gary Bending⁵ | Sally Hilton⁵ | David Bass^{2,6} | Fabien Burki¹

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



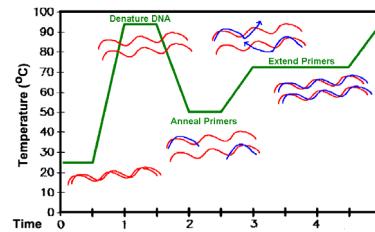
PCR

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



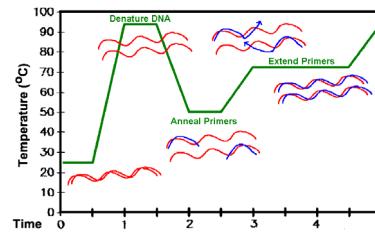
PCR

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



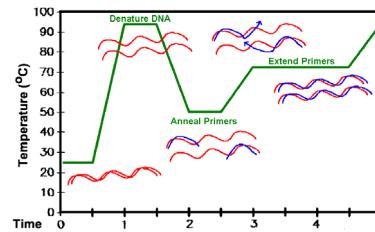
PCR

- 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!



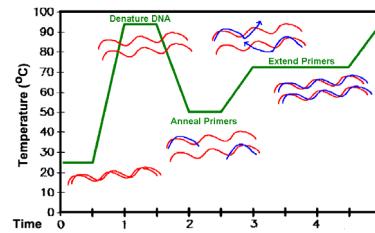
PCR

- 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!**

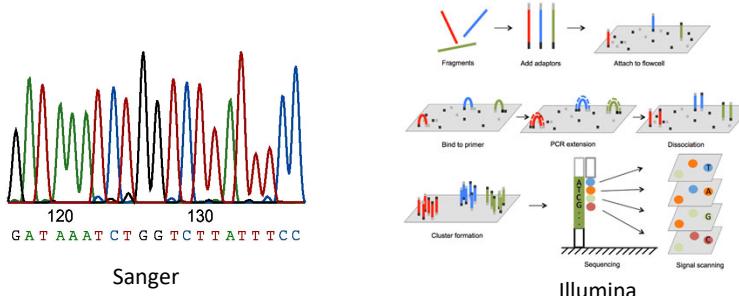


Sample types

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

PCR-induced errors

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



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

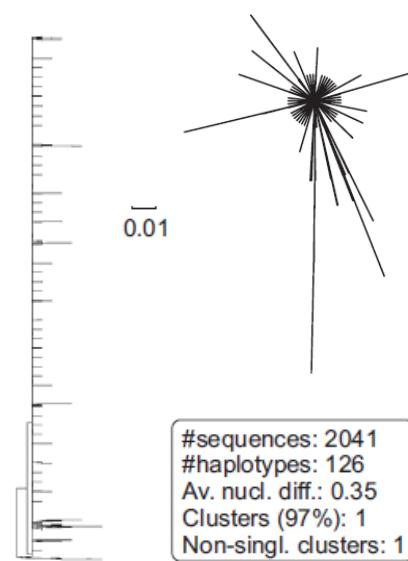
Daniel L. Lindner¹, Tor Carter², R. Henrik Nilsson³, Marie Davey^{2,4}, Trond Schumacher² & Howard K. Burge¹

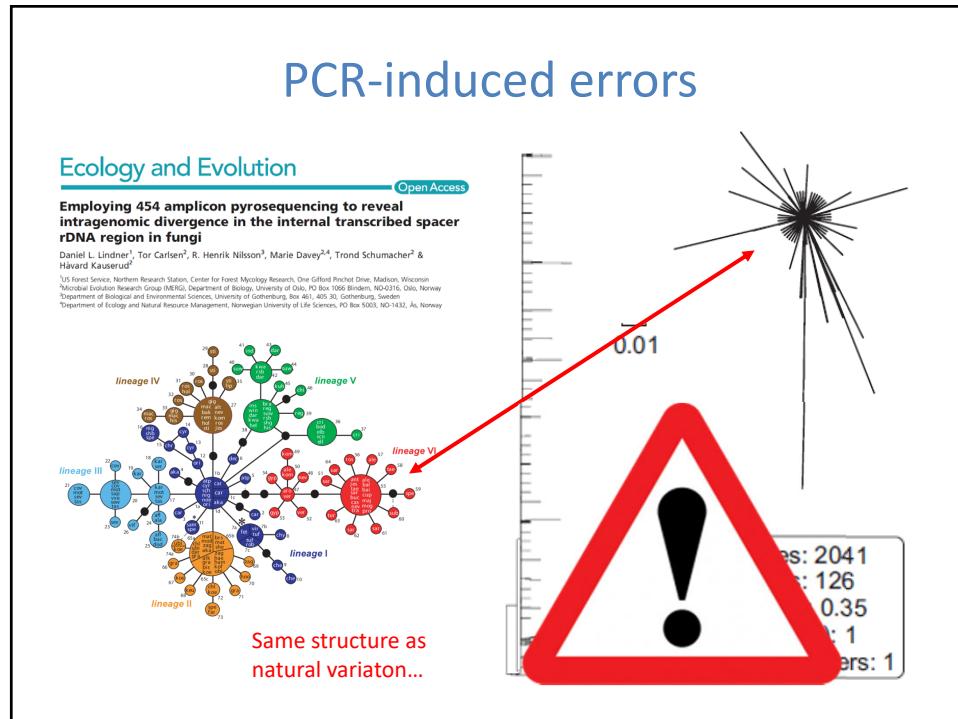
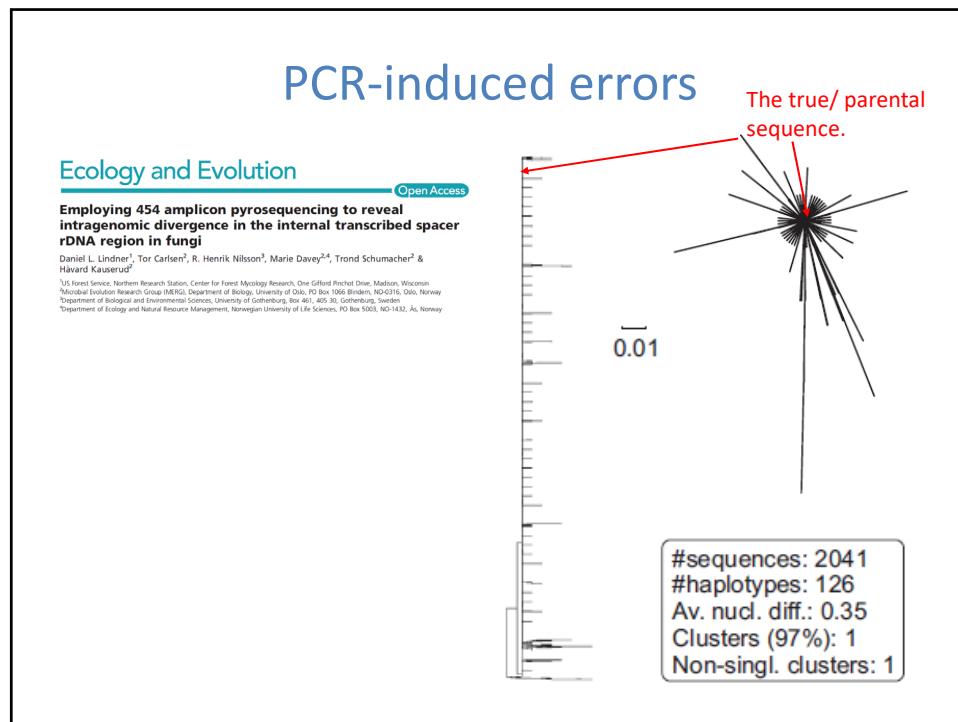
¹US Forest Service, Northern Research Station, Center for Forest Mycology Research, One Gifford Pinchot Drive, Madison, Wisconsin

²Microbial Evolution Research Group (MERG), Department of Biology, University of Oslo, PO Box 1066 Blindern, NO-0316, Oslo, Norway

³Department of Biological and Environmental Sciences, University of Gothenburg, Box 460, 405 30, Gothenburg, Sweden

⁴Department of Ecology and Natural Resource Management, Norwegian University of Life Sciences, PO Box 5300, NO-1432, Ås, Norway





PCR-induced errors

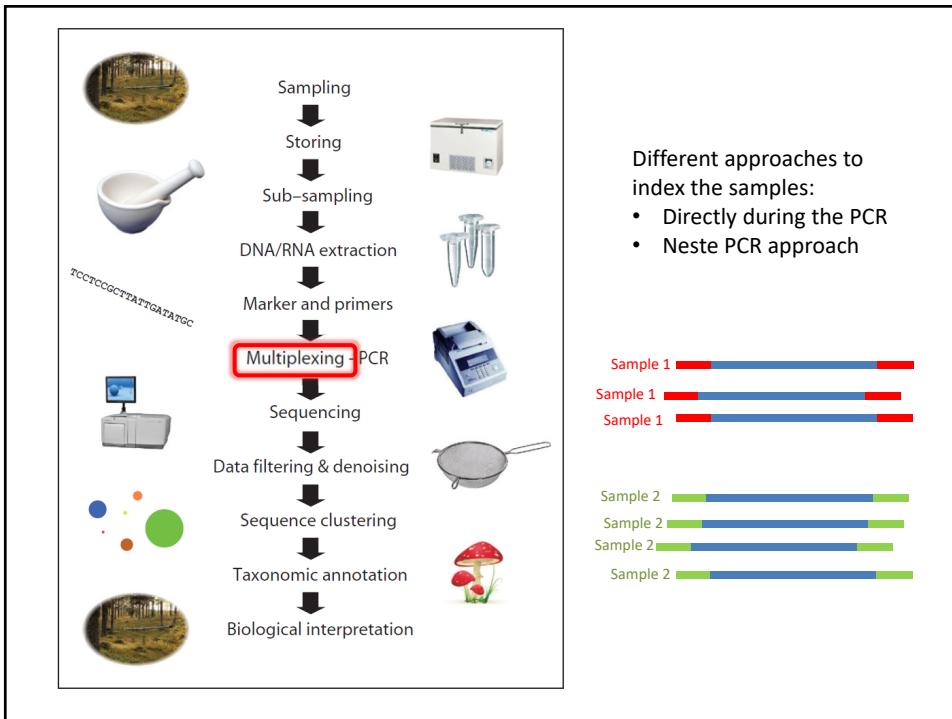
- Chimeric sequences

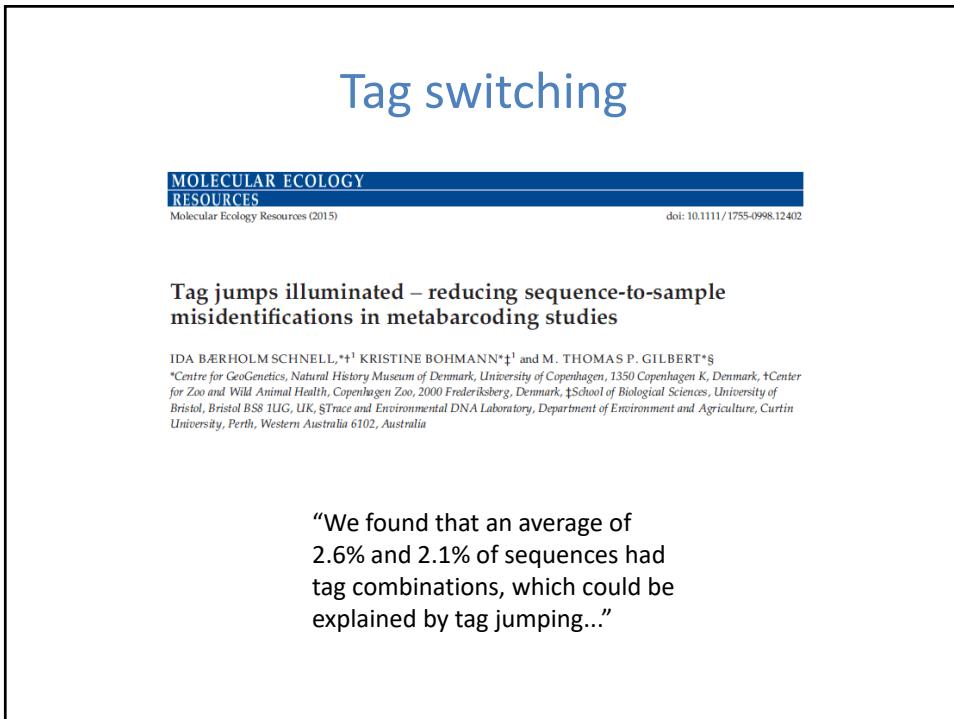
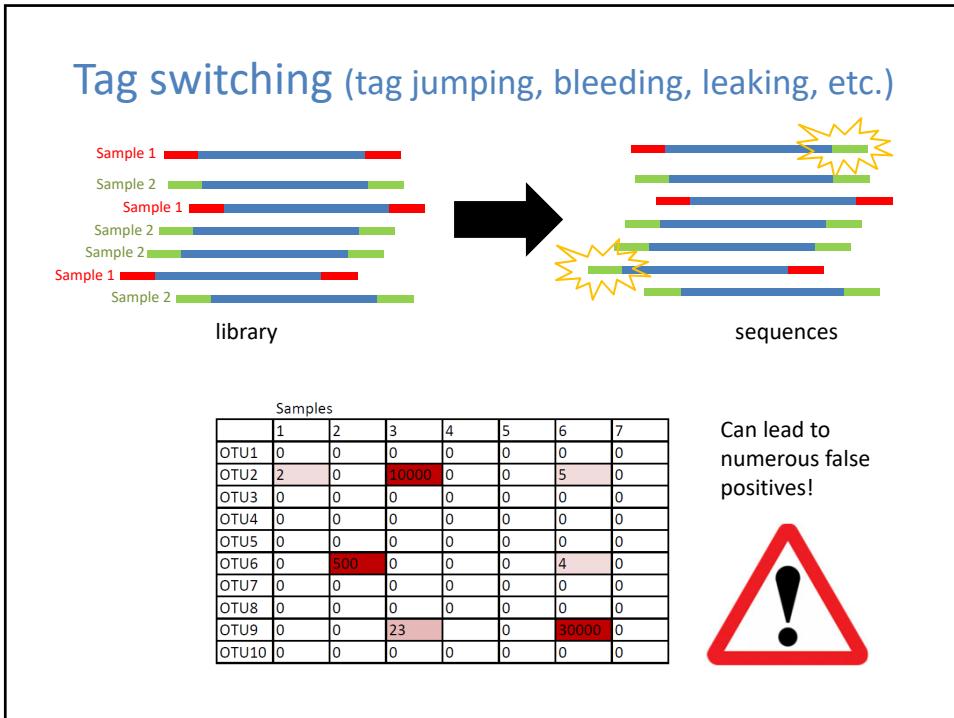
The level of chimeric sequences depends on how variable the marker is!

Can reduce the problem with certain PCR settings

MOLECULAR ECOLOGY RESOURCES
Molecular Ecology Resources (2016)
doi: 10.1111/1755-0998.12022

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 (Engen), Department of Biosciences, University of Oslo, P.O. Box 1066 Blindern, NO-0316 Oslo, Norway





Tag switching

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

Sample types

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

EDITORIAL

MOLECULAR ECOLOGY WILEY

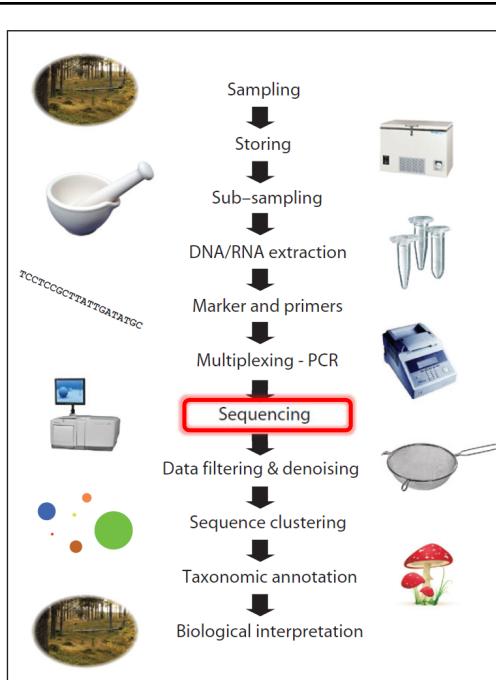
DNA metabarcoding—Need for robust experimental designs to draw sound ecological conclusions

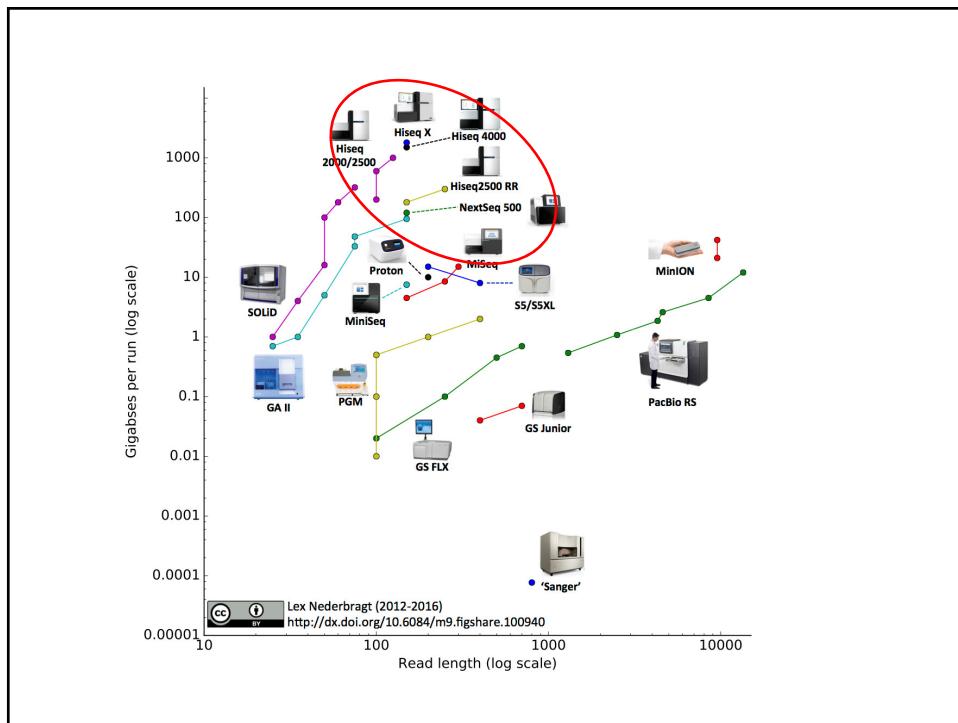
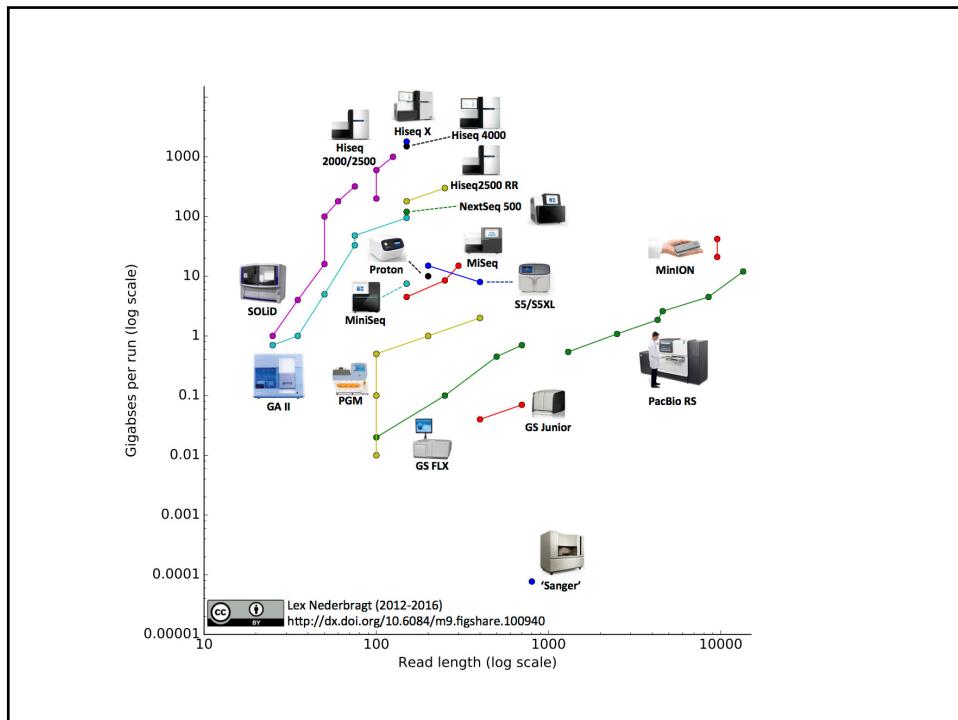
Zinger et al. 2019. Molecular Ecology Resources

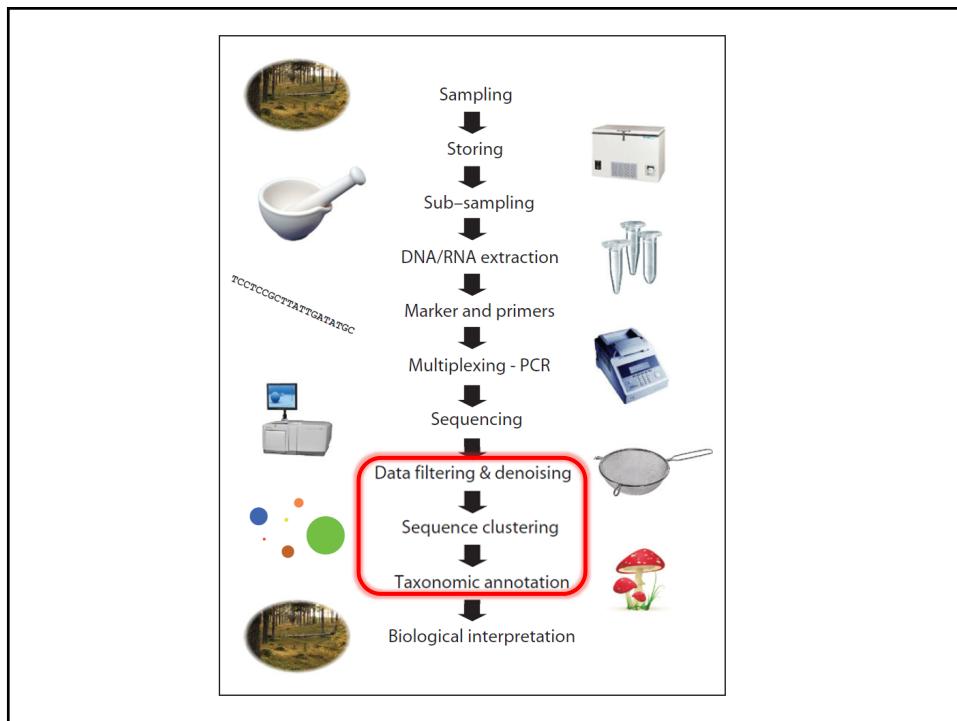
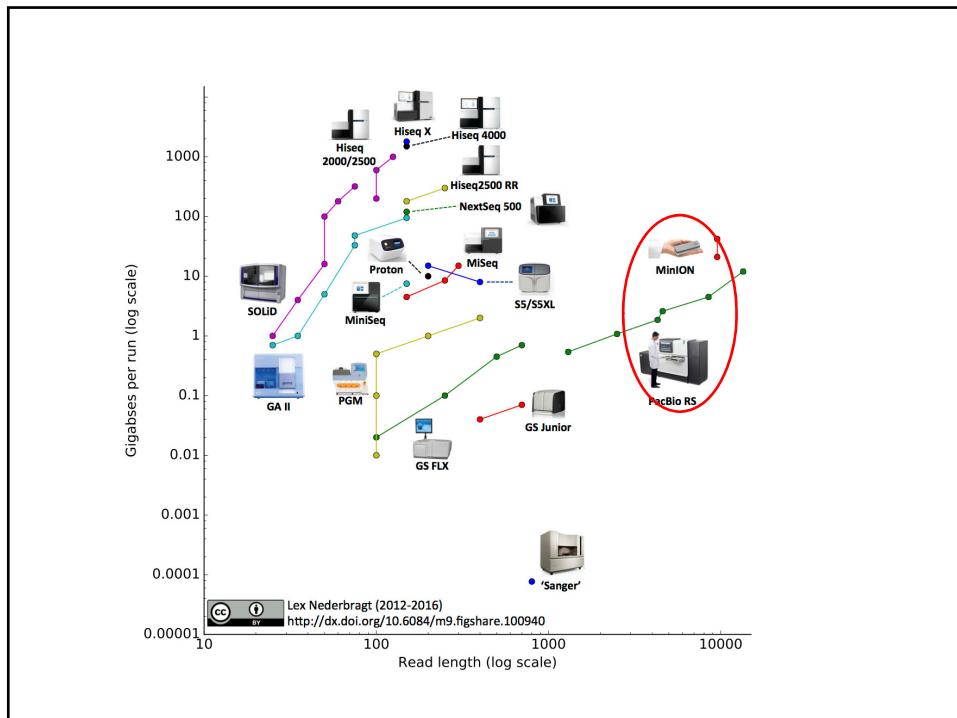
Tag switching

- The problem can be reduced or controlled for by:
 - Tagging in both ends with unique tag combinations
 - Rinse the PCR amplicons thoroughly
 - Avoid PCR step during the final library preparations steps before sequencing (i.e. when adaptors are introduced)
 - Include positive controls during PCR (mock community) → can better identify the level of switching/leakage
 - (Drastic) *ad hoc* solution: Process your OTU/sample matrix by removing low frequency occurrences (proportion-wise)

	Samples						
	1	2	3	4	5	6	7
OTU1	0	0	0	0	0	0	0
OTU2	2	0	0	100000	0	0	5
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	0	90000	0
OTU10	0	0	0	0	0	0	0







Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

Cleaning of tag bleeding

OTU modifications

Positive negatives

Singleton removal

Data transformation (e.g.
rarification)

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

Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

Cleaning of tag bleeding

OTU modifications

Positive negatives

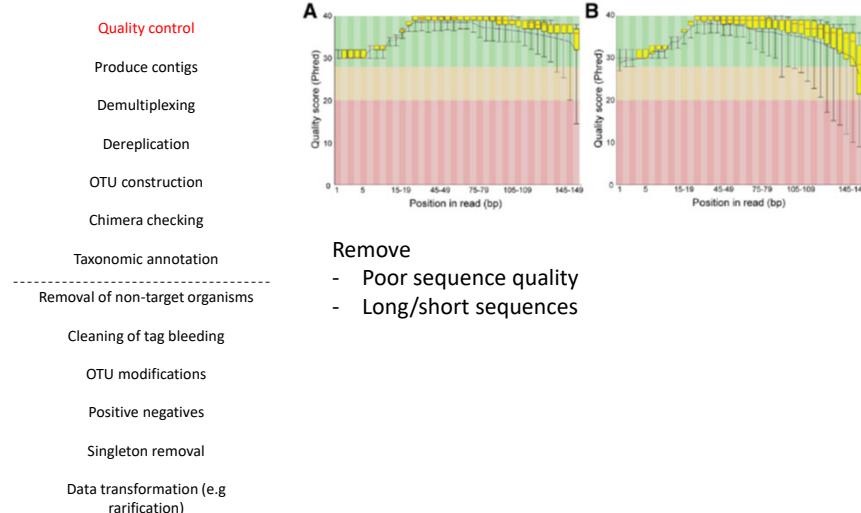
Singleton removal

Data transformation (e.g.
rarification)

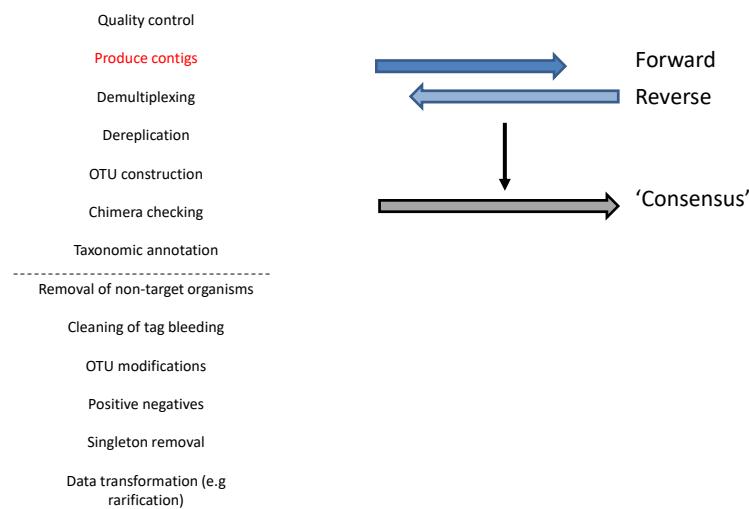
Bioinformatics steps

Post processing steps

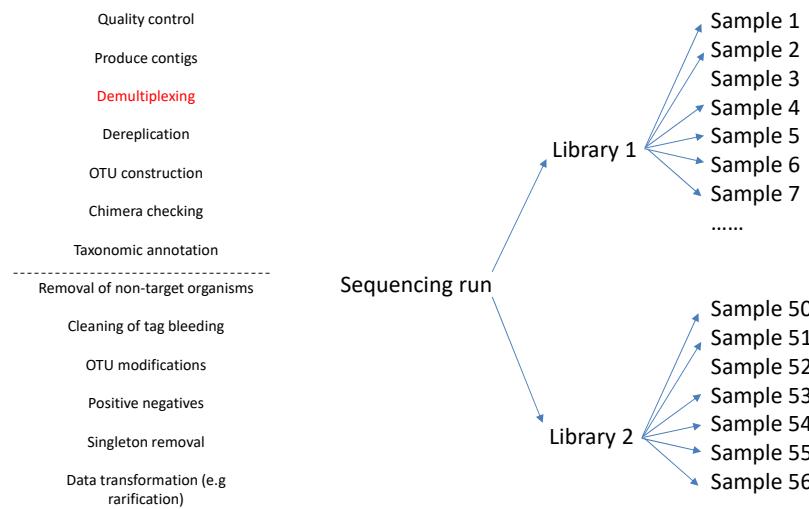
Bioinformatics – main steps



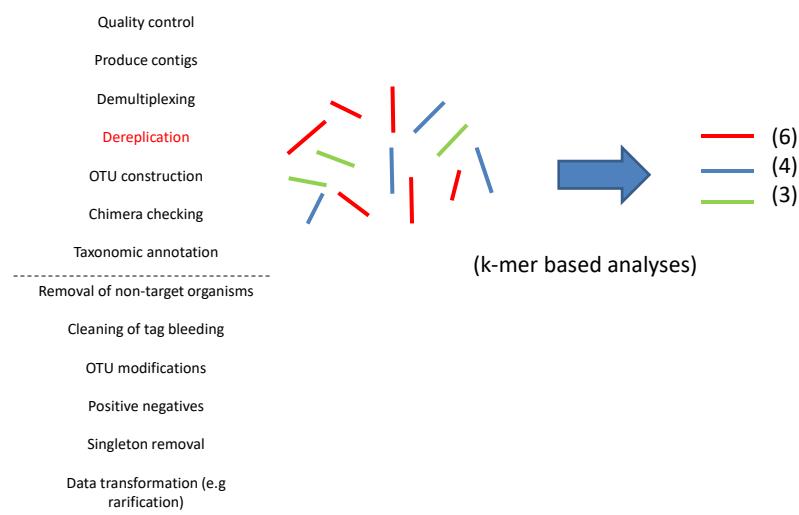
Bioinformatics – main steps



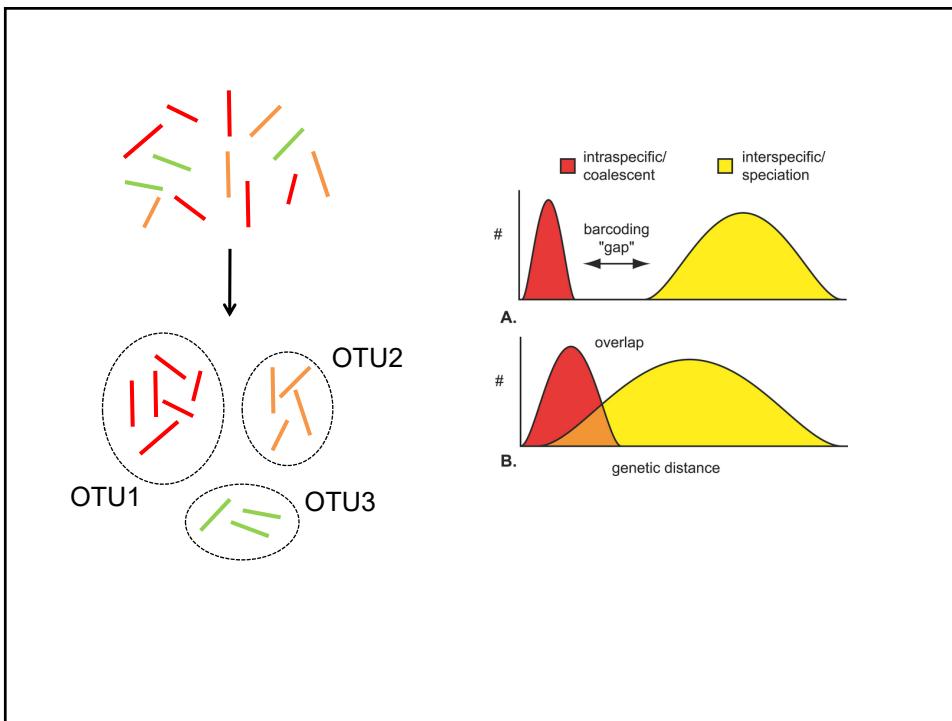
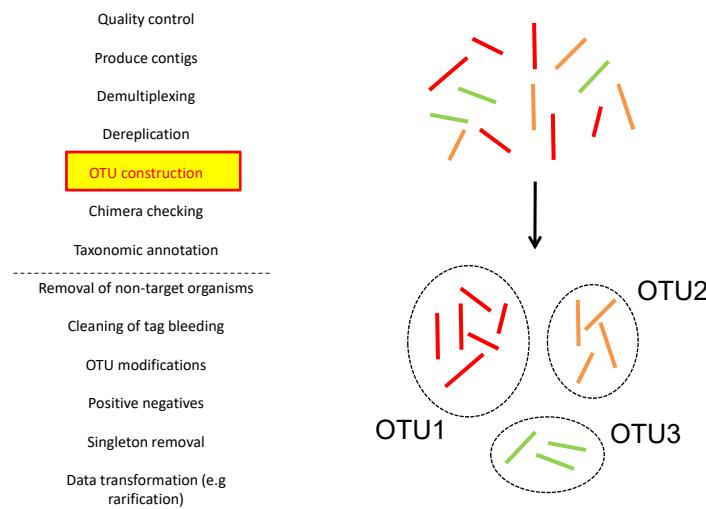
Bioinformatics – main steps

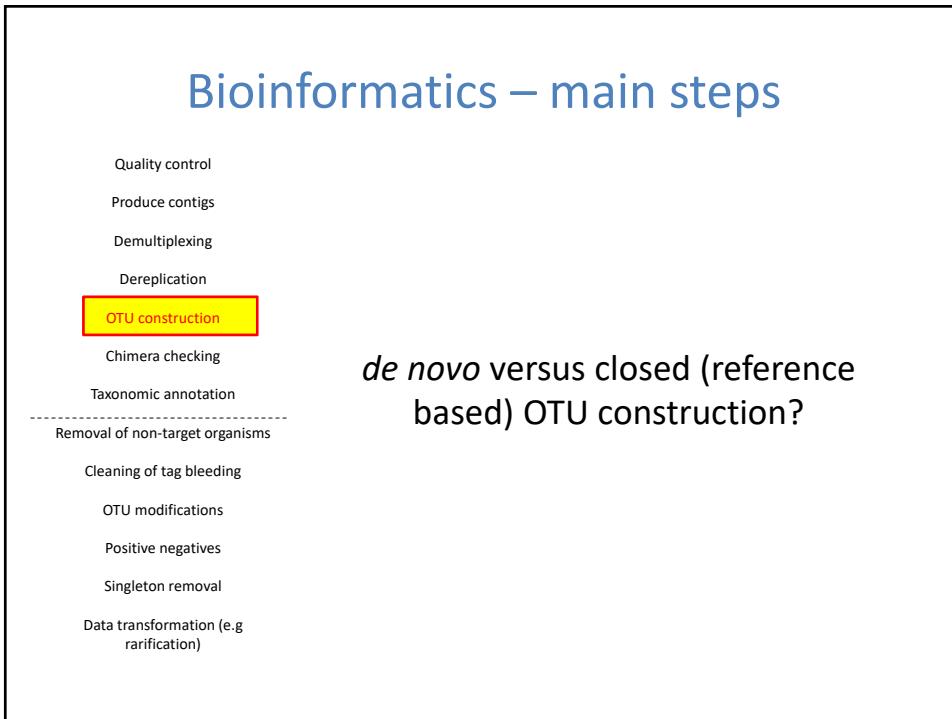
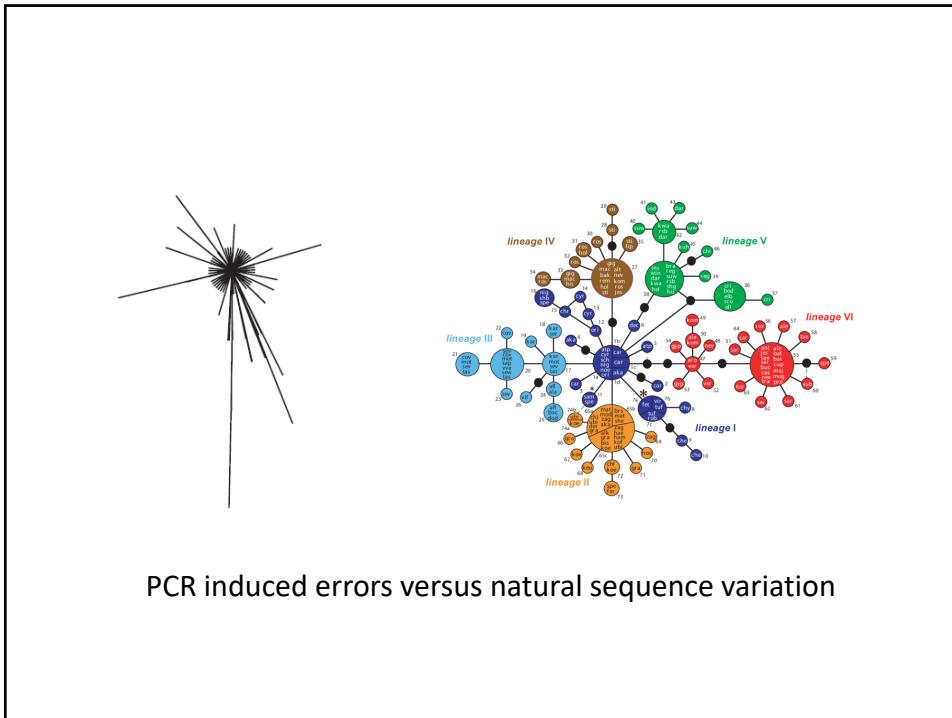


Bioinformatics – main steps



Bioinformatics – main steps





Bioinformatics – main steps

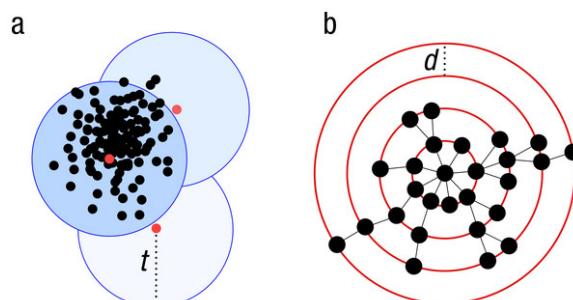
- Quality control
- Produce contigs
- Demultiplexing
- Dereplication
- OTU construction**
- Chimera checking
- Taxonomic annotation
-
- Removal of non-target organisms
- Cleaning of tag bleeding
- OTU modifications
- Positive negatives
- Singleton removal
- Data transformation (e.g. rarification)

de novo versus closed (reference based) OTU construction

Bioinformatics – main steps

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

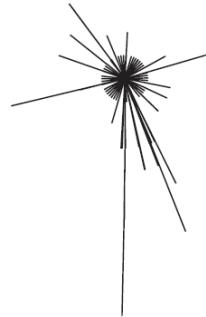
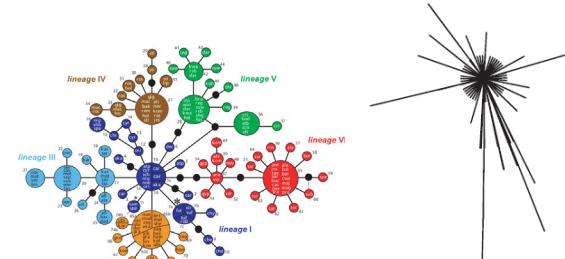
Many different clustering approaches



Mahé et al. 2014

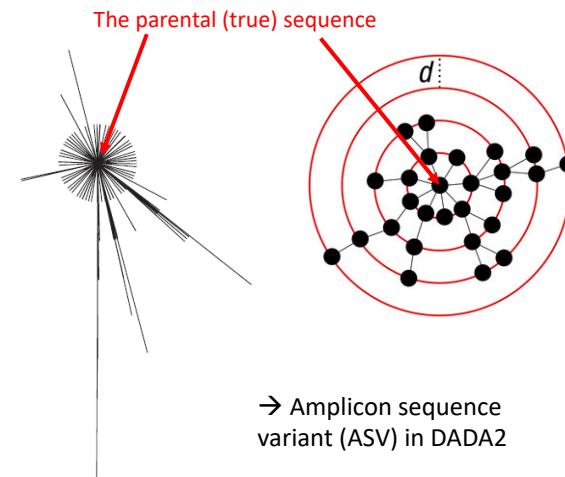
Bioinformatics – main steps

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

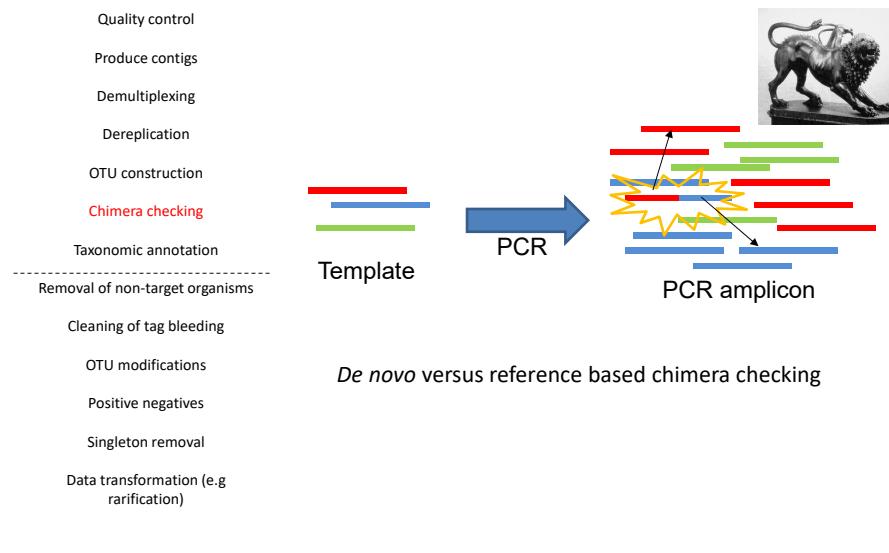


Bioinformatics – main steps

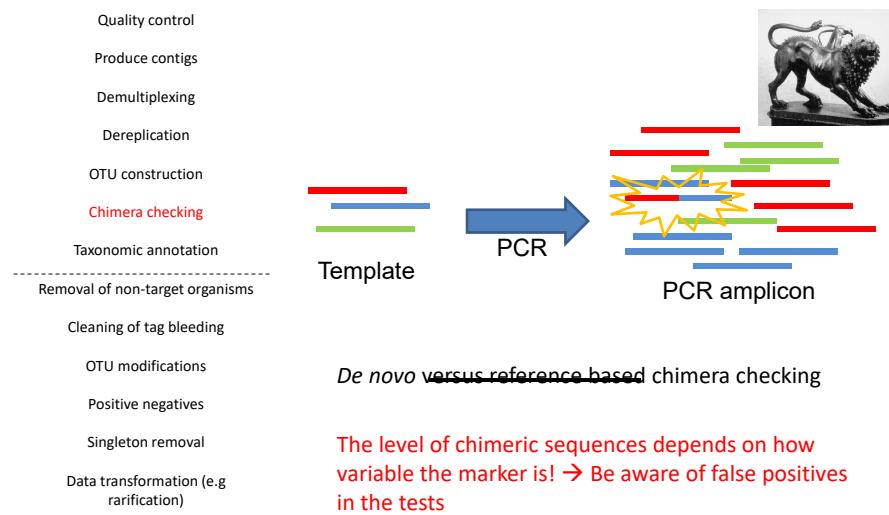
- Quality control
- Produce contigs
- Demultiplexing
- Dereplication
- OTU construction**
- Chimera checking
- Taxonomic annotation
-
- Removal of non-target organisms
- Cleaning of tag bleeding
- OTU modifications
- Positive negatives
- Singleton removal
- Data transformation (e.g. rarification)



Bioinformatics – main steps



Bioinformatics – main steps



Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

Cleaning of tag bleeding

OTU modifications

Positive negatives

Singleton removal

Data transformation (e.g.
rarification)



Simple matching → probabilistic assignment

Bioinformatics – main steps

Quality control

Produce contigs

Demultiplexing

Dereplication

OTU construction

Chimera checking

Taxonomic annotation

Removal of non-target organisms

Cleaning of tag bleeding

OTU modifications

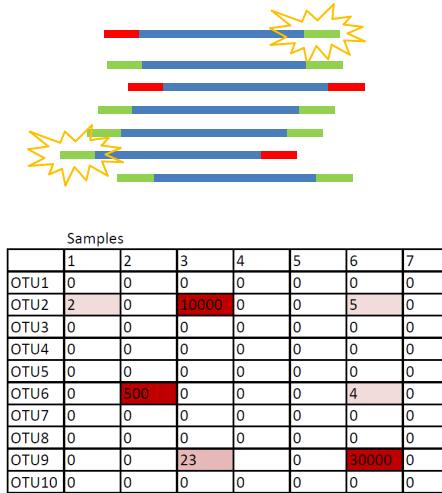
Positive negatives

Singleton removal

Data transformation (e.g.
rarification)

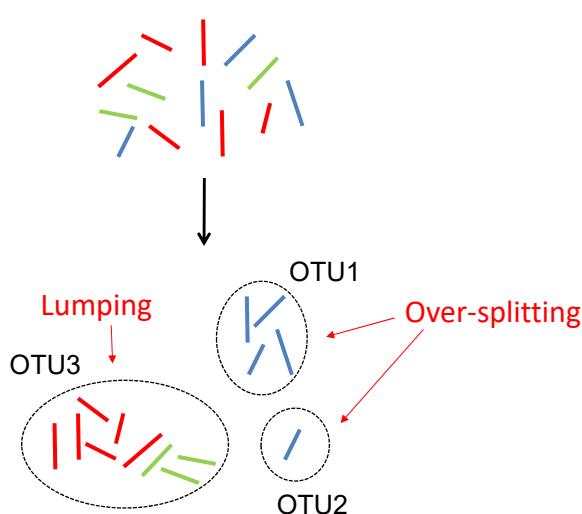
Bioinformatics – main steps

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

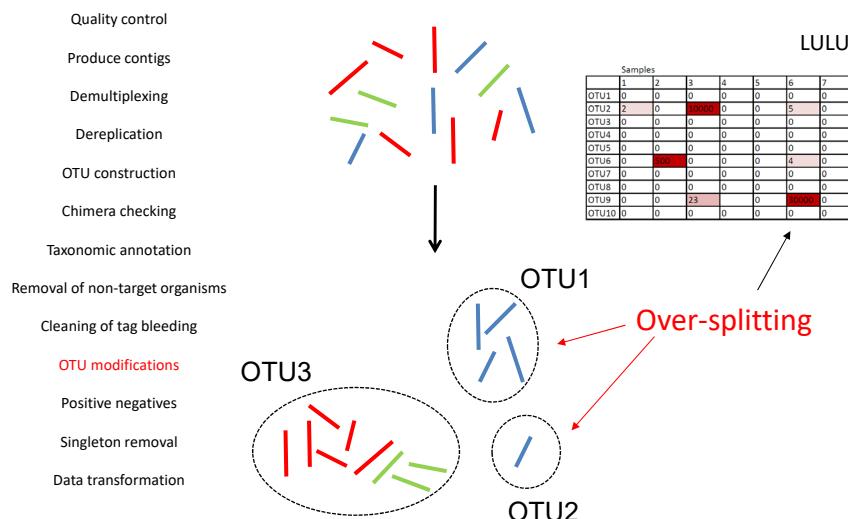


Bioinformatics – main steps

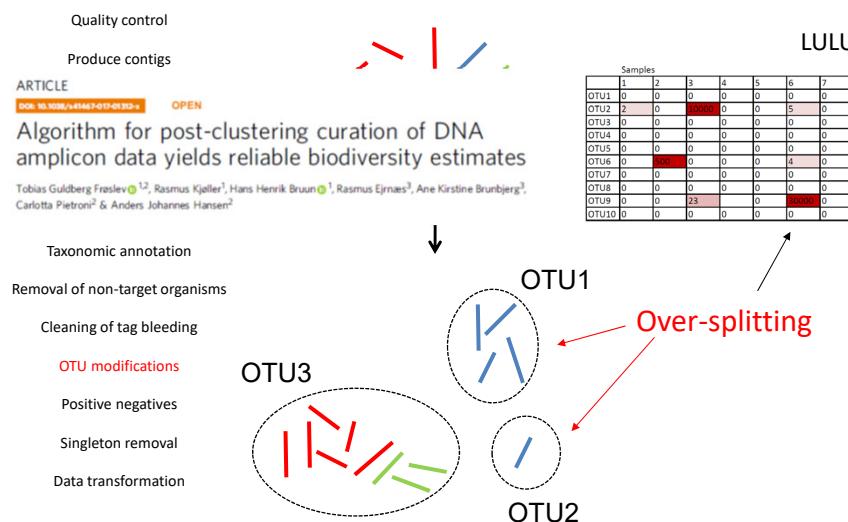
Quality control
 Produce contigs
 Demultiplexing
 Dereplication
 OTU construction
 Chimera checking
 Taxonomic annotation
 Removal of non-target organisms
 Cleaning of tag bleeding
OTU modifications
 Positive negatives
 Singleton removal
 Data transformation



Bioinformatics – main steps



Bioinformatics – main steps



Bioinformatics – main steps

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

External or internal contaminants?



RESEARCH ARTICLE
Novel Systems Biology Techniques
OPEN ACCESS

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

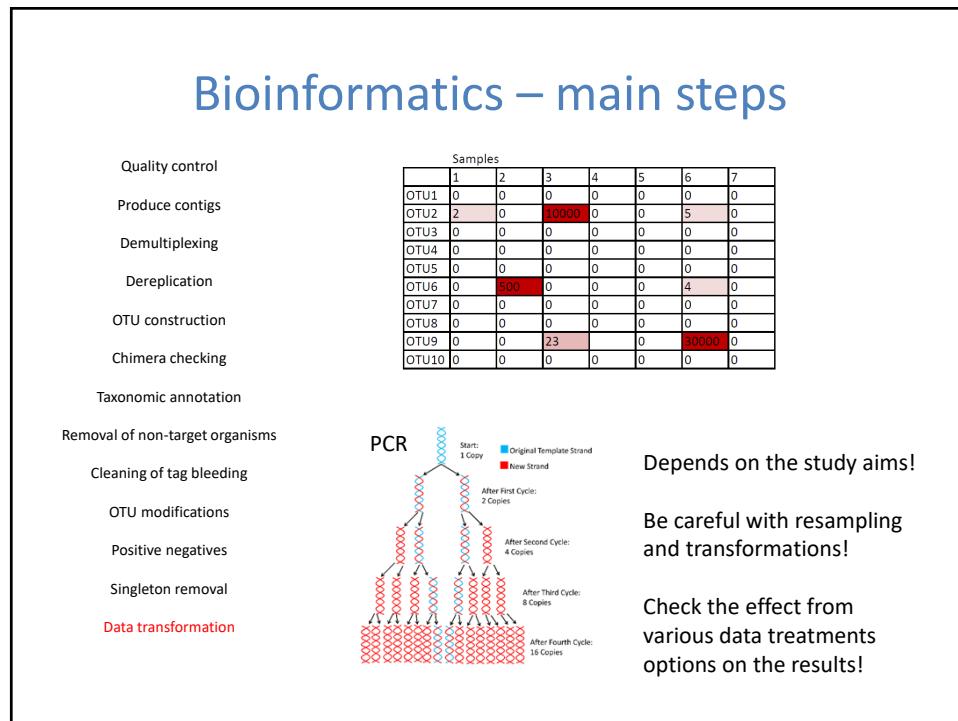
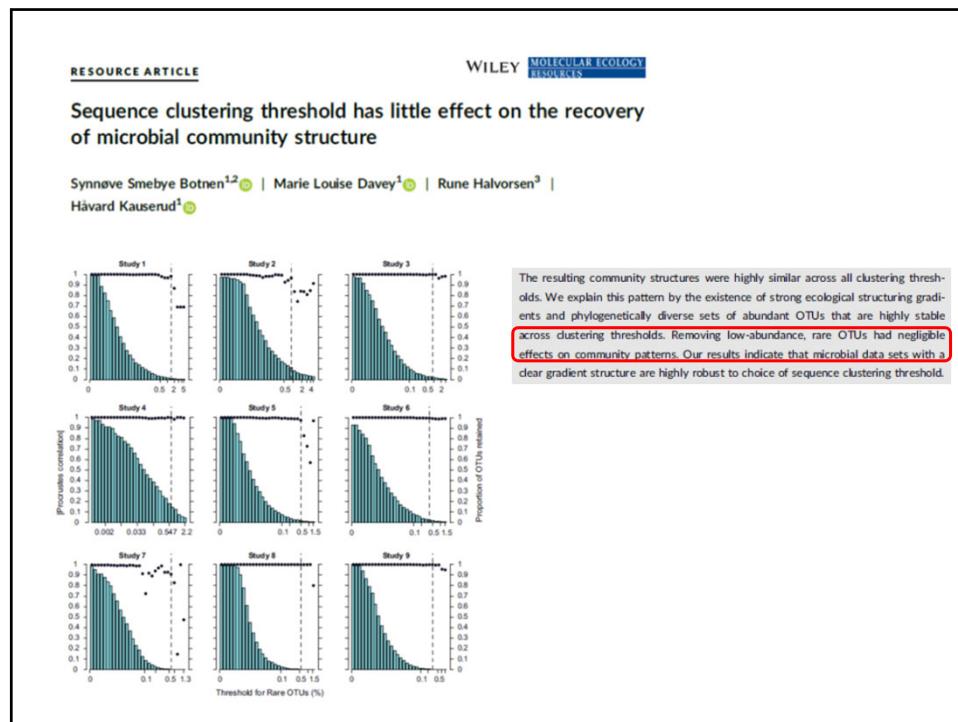
Jeremiah J. Minich,^{a,*} Jon G. Sanders,^{b,*} Amnon Amir,^b Greg Humphrey,^b Jack A. Gilbert,^{c,d} Rob Knight^{b,c,e,f}

Should be very careful how you treat positive negatives!!

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



What is a ‘singleton’?
Depends on your sequencing depth..



The importance of controls

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

Conclusions

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

INVITED REVIEWS AND SYNTHESSES

WILEY MOLECULAR ECOLOGY

Environmental DNA metabarcoding: Transforming how we survey animal and plant communities

Kristy Deiner¹ | Holly M. Bik² | Elvira Mächler^{3,4} | Mathew Seymour⁵ | Anais Lacoursière-Roussel⁶ | Florian Altermatt^{7,8} | Simon Creer⁵ | Ilana Bista^{5,7} | David M. Lodge¹ | Natasha de Vere^{8,9} | Michael E. Pfrender¹⁰ | Louis Bernatchez⁶

EDITORIAL

MOLECULAR ECOLOGY WILEY

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

Zinger et al. 2019. Molecular Ecology Resources